CALIBRATED PROBABILISTIC QUANTITATIVE PRECIPITATION FORECASTS BASED ON THE MRF ENSEMBLE

THESIS

Frederick Anthony Eckel, Captain, USAF

AFIT/GM/ENP/98M-02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GM/ENP/98M-02

CALIBRATED PROBABILISTIC QUANTITATIVE PRECIPITATION FORECASTS BASED ON THE MRF ENSEMBLE

THESIS

Frederick Anthony Eckel, Captain, USAF

AFIT/GM/ENP/98M-02

AFIT/GM/ENP/98M-02

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Approved for public release; distribution unlimited

AFIT/GM/ENP/98M-02

CALIBRATED PROBABILISTIC QUANTITATIVE PRECIPITATION FORECASTS BASED ON THE MRF ENSEMBLE

THESIS

Presented to the Faculty of the Graduate School of Engineering
of the Air Force Institute of Technology
Air University
Air Education and Training Command
In Partial Fulfillment of the Requirements for the
Degree of Masters of Science in Meteorology

Frederick Anthony Eckel, B.S.
Captain, USAF

March 1998

AFIT/GM/ENP/98M-02

CALIBRATED PROBABILISTIC QUANTITATIVE PRECIPITATION FORECASTS BASED ON THE MRF ENSEMBLE

Frederick Anthony Eckel, B.S.
Captain, USAF

Approved:

Michael K. Walters, LT COL, USAF
Advisory Committee Chairman

4 MAR 98
Date

Clifton E. Dungey, MAJ, USAF
Advisory Committee Member

4 MAR 98
Date

Daniel E. Reynolds
Advisory Committee Member

4 March 1998
Date

David E. Weeks
Advisory Committee Member

4 March 98
Date

Acknowledgments

Because this research was so novel and interdisciplinary, I sought and received assistance from many individuals to whom I owe a great deal of thanks. First of all, thank you to Jeff Doran for suggesting the topic of this research and for connecting me up with the right sources. I would like to thank my advisor, LtC Mike Walters, for giving me so much latitude in this project and for keeping me focussed on the goal.

Next on the list is Dan Reynolds. I really enjoyed the countless hours we spent figuring out the nature of ensemble forecasting. I could not have understood this material without his insight and enthusiasm.

I am forever indebted to Tom Hamill. He was always quick and thorough in responding to my barrage of questions concerning his work. His numerous suggestions during the research and his review of the draft thesis were invaluable.

Another person to whom I owe many thanks is Zoltan Toth for his collaboration on this project. Using his connections and expertise in this subject, he was willing and able to provide assistance which kept this project rolling.

Two individuals were instrumental to this project in providing the massive amounts of data required as well as in consulting on its handling. Mike Baldwin saved me an incredible amount of time by analyzing the observed cumulative precipitation data in the most useful format. Andrew Loughe organized and provided the largest part of the data, the archived ensemble forecasts.

Lastly, I would like to thank my stepmother, Jean Eckel, for her tremendous efforts in editing this document.

Table of Contents

List of Figures

List of Tables

Table Page

Abstract

Probabilistic quantitative precipitation forecasts (PQPF) based on the medium range forecast (MRF) ensemble are currently in operational use below their full potential quality (i.e., accuracy and reliability). This unfulfilled potential is due to the MRF ensemble being adversely affected by systematic errors which arise from an imperfect model and less than ideal ensemble initial perturbations. This thesis sought to construct a calibration to account for these systematic errors and thus produce higher quality PQPF. Systematic errors were explored with the use of the verification rank histogram, which tracks the performance of the ensemble. The information in these histograms was then used in interpreting MRF ensemble forecasts to produce calibrated PQPF.

While the calibration technique did noticeably improve the quality of PQPF, its usefulness was bounded by the natural predictability limits of cumulative precipitation. It was discovered that higher levels of cumulative precipitation cannot be reliably predicted in the medium range.

This limitation is likely due to the extremely high spatial and temporal variability of precipitation. Due to this limit of predictability, for significant levels of precipitation (high threshold), the calibration designed in this thesis was found to be useful only for short range PQPF. For low precipitation thresholds, the calibrated PQPF did prove to be of value in the medium range.

CALIBRATED PROBABILISTIC QUANTITATIVE PRECIPITATION FORECASTS BASED ON THE MRF ENSEMBLE

1. Introduction

a. Background

The primary tool of today's weather forecaster is output from numerical weather prediction (NWP) models. Because of this reliance on numerical guidance for lead times beyond about six hours, improving the accuracy of NWP models represents the best opportunity for improving weather forecasting capabilities. Even a seasoned forecaster who correctly interprets the NWP prognostic charts in making a forecast for his local area is still at the mercy of the quality of the charts.

Efforts over the past 50 years to improve NWP have been remarkably successful. As computer power increased, newer atmospheric models included more realistic physics on finer grid scales while integrating with smaller time steps. This has brought about a steady improvement in the forecast quality of NWP models (Mullin, 1993). However, continuing to focus on improving NWP through increasing the resolution of a single forecast may not be the best method to pursue for future advancement of NWP (Brooks and Doswell, 1993; Lorenz, 1993). A relatively new technique, called ensemble weather forecasting, appears to offer the next wave of improvement for NWP.

Ensemble weather forecasting is a concept which merges NWP and chaos theory. In the standard application of a NWP model, the model is run once from the initial state

(called the analysis) out to some future time. The analysis is the best estimate of the true state of the atmosphere. Using a NWP model, a single deterministic forecast state of the atmosphere at any desired time into the future can be generated. The problem is that there are errors in the analysis which make any forecast state of the atmosphere in error as well, even if the model is perfect. What makes matters worse for weather forecasting is that because of chaos, these errors often increase rapidly as the time span of the forecast increases. The sources of errors in the analysis include lack of both accuracy and precision in observations, lack of complete spatial coverage of observations, and errors in fitting observations to gridded fields (Toth and Kalnay, 1993).

Ensemble weather forecasting attempts to gain more information from the NWP process by creating a spread of possible forecasts versus a single forecast created with standard NWP forecasting. The basic method is to run an atmospheric model *n* times with each ensemble run started from a slightly different, or *perturbed*, initial state. This results in *n* forecast states of the atmosphere at the desired forecast time, but of course only one true state. These *n* forecasts are all in error but will likely encompass the truth if the ensemble method is correctly applied. Each of the *n* unique forecasts is termed a *member* of the ensemble.

There are four major applications for an ensemble of model output (Anderson, 1996; Toth and Kalnay, 1993; Tracton and Kalnay, 1993). First, ensemble forecasts can give an idea of confidence in the single, standard model run made from the analysis. The wider the ensemble spread at some forecast time, the more likely it is that any one model run, including the standard run, will be further from the truth. Secondly, averaging the *n*

erred forecasts together should produce a forecast closer to the true atmosphere than the one standard run. Thirdly, ensemble forecasts can be examined for *clusters* of similar solutions to narrow down the most likely forecast scenario(s). Lastly, since the *n* forecasts give *n* possible values for any meteorological parameter, probability forecasts can be produced for that parameter. This last application is the focus of this research.

Probabilistic quantitative precipitation forecasts (PQPF) are predictions of the likelihood of exceeding a threshold of cumulative precipitation at some location in a given amount of time. For example, in the 48 – 72 hour forecast valid period, the chance of precipitation > 5.0 mm at Andrews AFB may be 75%. Normally, such a forecast would be produced by a weather forecaster using standard NWP output, climatological data, and/or a variety of other tools. As will be described in detail later, an ensemble forecast of cumulative precipitation can also be used to derive PQPF. In general, the more members of an ensemble which forecast an amount higher than the threshold, the more likely it is that the verification (observed amount) will also exceed the threshold.

Quality of probabilistic forecasts is measured in two different ways, namely reliability and accuracy. Reliability is determined by how well the forecast probability represents the observed occurrence of the event over many samples of forecasts and corresponding observations. Continuing the above example, suppose that over a period of two years there were 131 cases when the chance of precipitation > 5.0 mm was forecast at 75%. Of these 131 cases, precipitation > 5.0 mm was observed for 96. This means that the forecasts were highly reliable with an observed 73% occurrence of precipitation > 5.0 mm for the sample space of all the cases.

The other measure for probabilistic forecasts, quite different from reliability, is accuracy. Accuracy is a measure, usually a mean square error, of the difference between the forecast probability and the occurrence or nonoccurrence of the event. In the above example for a single 75% forecast, suppose that precipitation > 5.0 mm is observed in the valid period. While this was a good forecast, a more accurate forecast would have been anything higher than 75%.

b. Problem and Objective

PQPF derived from the medium range forecast (MRF) ensemble, produced daily at the National Centers for Environmental Prediction (NCEP), display skill but are adversely affected by systematic errors. These errors are a result of both imperfections in the atmospheric model and inadequacies in design of the ensemble. Since NCEP's PQPF production lacks any compensation for these systematic errors, PQPF are currently being used at a level below their full potential quality.

The goal of this research was to produce calibrated PQPF from MRF ensemble precipitation forecasts which compensate for the systematic errors. With a good calibration technique, medium range PQPF with improved reliability and higher accuracy may be possible.

c. Importance of Research

This research has both a general and a specific importance to the United States Air Force. The general importance involves the overall application of ensemble forecasting, while the specific importance is tied to the objective of this research.

While the benefits of medium range (2 days < forecast time < 2 weeks) ensemble forecasting have been clearly demonstrated for many years now, the vast majority of weather forecasters in the USAF have never even heard of the term. This is most likely due to the fact that weather operations in the Air Force concentrate on short range forecasts (forecast time < 2 days) in supporting the flying mission. Short range ensemble forecasting (SREF) has shown promise but is still under development (Hamill and Colucci, 1997).

Since there are still many questions concerning SREF, it is difficult to say how much improvement could be realized by incorporating ensemble techniques into everyday weather forecasting in the USAF. It is the firm belief of the author that the potential benefits would be significant. The general importance of this thesis therefore is to illustrate for the Air Force Weather Agency the potential of ensemble forecasting by clearly describing the technique and by demonstrating just one of its many applications.

The specific value of this research comes from the obvious benefits of higher quality PQPF. Precipitation is a weather parameter which has a large impact on many aspects of USAF operations. It adversely affects air terminal operations, ground operations, weapons targeting, reconnaissance operations, communication, and radar, just to name a few. An improvement in medium range precipitation forecasts would greatly benefit planning of such operations, thus increasing mission effectiveness.

d. Summary of Key Results

The most important result of this research is that the calibrated PQPF produced by this project showed a significant improvement in both accuracy and reliability over the

current uncalibrated PQPF. This demonstrated that the systematic errors of the MRF ensemble can be accounted for in making probabilistic forecasts of cumulative precipitation. What is even more promising is that the technique used in this research could theoretically be applied to any weather parameter.

Another important finding was that higher levels of cumulative precipitation can not be reliably forecast in the medium range, with or without a calibration. Even at the lowest threshold, the calibrated PQPF are only of value out to about six days. Beyond this point, a climatologically based PQPF is the most reliable.

The third key result of this research was that systematic errors were found to be stationary but with a regime dependence. This means that errors between the atmosphere and the ensemble reoccur for a given atmospheric scenario. Because of this, the best calibration to the ensemble should be based on many specific aspects of the event being forecast (i.e., geographical location, season, atmospheric stability, etc.).

e. Thesis Organization

In this chapter, the general background of ensemble weather forecasting has been introduced, followed by the main points of this research including its importance, the problem statement, and its key results. Chapter 2 will cover the detailed background of chaos theory and ensemble forecasting. The research methodology will be described in chapter 3, followed by the research findings in chapter 4. Lastly, chapter 5 will give the conclusions of the findings, recommendations, and possible future research.

2. *Theoretical Background*

a. Overview

This chapter covers the theoretical basis of the methods used in this research, namely chaos theory and ensemble weather forecasting. These are two extremely diverse subject areas which can not possibly be thoroughly described within the confines of a Masters thesis. The following discussion is limited to the main ideas of these subjects and details which are involved in this research.

b. Chaos Theory

While ensemble weather forecasting is an expansion of the NWP process, chaos theory is the impetus for the whole ensemble concept. Knowledge of the basic principles of chaos theory is essential for understanding the methods and goals of ensemble forecasting. This chapter will therefore include a brief review of applicable elements of chaos theory.

Chaos theory describes the behavior and predictability of *dynamical systems*. The key idea of a dynamical system is that predictions of the system are possible based on the system's known initial state and some set of rules (Tsonis and Elsner, 1989). These rules are usually defined by an equation or set of equations. For predicting the atmospheric dynamical system, a typical model includes a set of differential equations with time as the independent variable. By integrating these equations, a prediction of the future state of the system can be made.

Before Edward Lorenz's ground breaking work in the 1960s, there were generally two schools of thought concerning prediction of complex dynamical systems such as the atmosphere (Lorenz, 1963; Mullin, 1993). One idea was that by somehow discovering and solving the governing differential equations of such a system, exact predictions out to any point into the future could be made, given that the initial and boundary conditions are known. The contrary opinion was that such systems are totally unpredictable after a certain time into the future because of completely random, indescribable interactions within the system.

Lorenz (1963) showed that both of these notions were in fact wrong. He discovered that there is indeed a limit to accurate predictions of future states of the atmosphere, but it is not due to random behavior. Furthermore, even if you do know the governing differential equations of a system, there is still a limit to accurate prediction. Lorenz presented the idea that the difficulty in predicting future states of a system is a result of sensitivity to the system's *initial conditions* (IC) and not a result of random behavior.

IC are the set of values for the dependent variables which completely describe the system at some starting point. Recall that the independent variable is time. Lorenz discovered that in a computer model of a chaotic dynamical system, changing the IC by a seemingly insignificant amount results in a very different set of solutions. For a short time into the forecast period, the two solutions evolve quite similarly. This makes it difficult to discern any difference between the solution begun from original IC vs. the altered IC. Further into the future, however, the two solutions diverge noticeably.

Continuing this scenario, suppose the original IC and subsequent evolution of solutions represent the true behavior of the system. The altered IC then represent the erred best estimate of the state of the system at the starting point. Predictions made by the model which initialized with the altered IC would be fairly accurate at first but become worthless at some point. In an attempt to predict the future, apparent randomness and unpredictability actually result from imperfect IC. Erred IC send the modeled evolution of the system off on a completely different path through *phase space* when compared to the true evolution.

Behavior of a dynamical system can be graphically represented in its phase space, where each dependent variable corresponds to a coordinate axis. A phase space path, or *trajectory*, is a continuous set of points which make up a curve. A single point on the trajectory gives the values of the system's dependent variables at some particular time, thus describing the state of the system. As time progresses, the direction the trajectory takes through phase space is determined by the dynamics of the system. To demonstrate this, consider the trajectory of an underdamped, simple harmonic oscillator. While this system is not actually chaotic (Lorenz, 1993), it is useful for visualizing several concepts of chaos.

This oscillator consists of a block attached to a spring and suspended from above (Figure 1a). Once displaced from equilibrium, the block will oscillate up and down many times before returning to the equilibrium position. For the purposes of this demonstration, consider the state of the system to be totally described by the block's

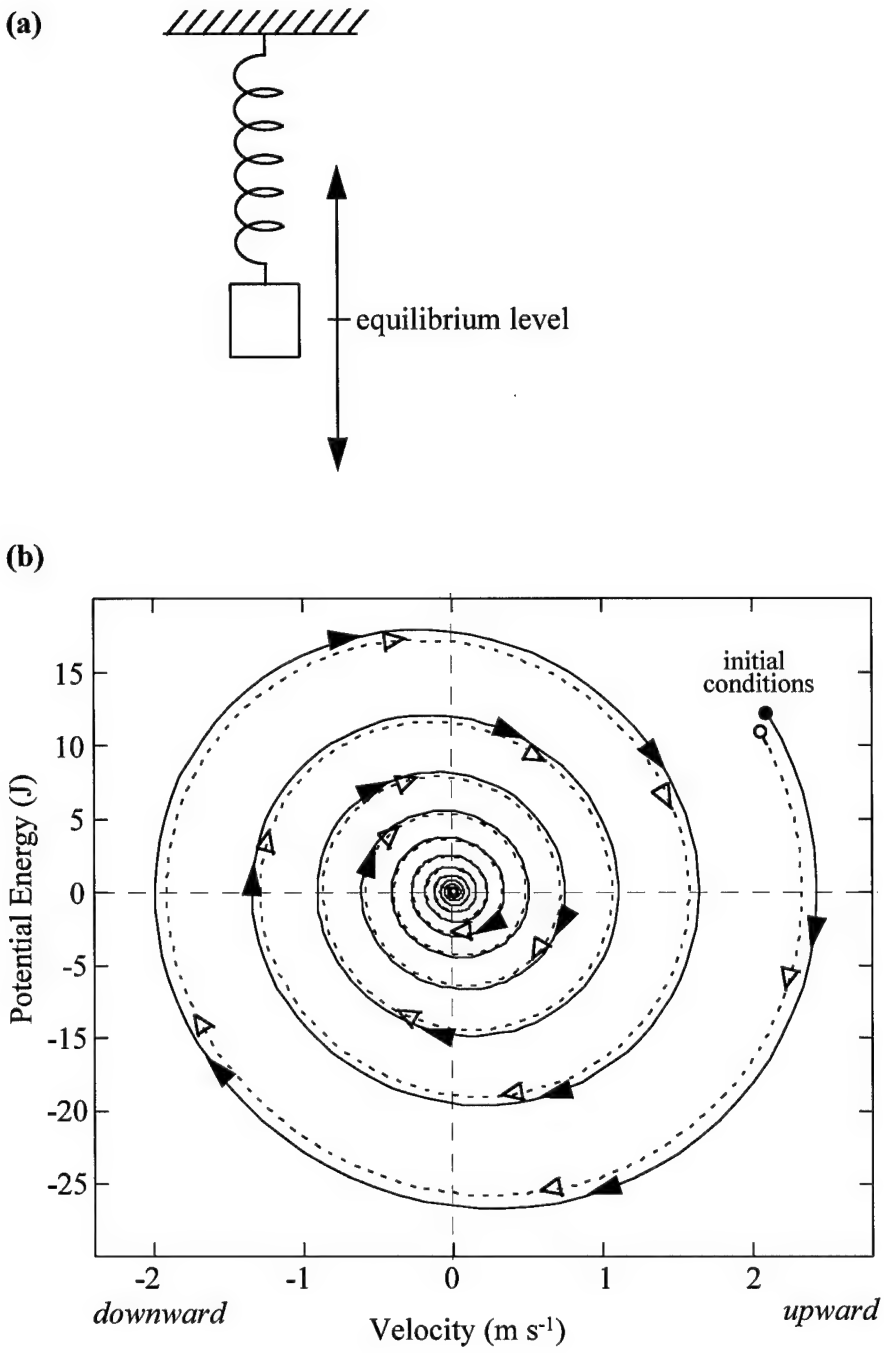


Figure 1. Underdamped simple harmonic motion. (a) Physical diagram of harmonic oscillator. The block is attached to a spring which is attached from above. The damping of the motion is provided by friction in the spring and friction between the block and the air. (b) Trajectory of simple harmonic oscillator in phase space. Arrows show the direction of time along the trajectory. The solid trajectory (with filled arrows) represents the true behavior of the system while the dashed line (with open arrows) represents an attempt to model and predict true behavior of the system.

velocity and its gravitational potential energy relative to the equilibrium position. At the initial time, the block is lifted then pushed downward, thus setting the IC for the system.

The IC and all points along a trajectory should be thought of as a vector in phase space extending from the origin to the point on the trajectory. In Figure 1b, the filled circle representing the true IC of the system is at the tip of vector (2.1 m s^{-1}, 12.2 J) which is not displayed. The solid curve shows the true trajectory of the dynamical system as time progresses. It is clear in the figure that the system's future position in phase space is totally deterministic. Each vector of velocity and potential energy values along the trajectory represents a unique state of the system at a particular time.

An important characteristic of a forced, dissipative chaotic system is that a trajectory never intersects itself (Lorenz, 1963). If the system were to return to a previous state, it would then repeat all the following states and get stuck on a periodic trajectory. If friction were completely removed from the example system, making it an undamped oscillator, its trajectory would then be an ellipse (a periodic trajectory). Friction is an example of a system parameter. Increasing friction in the spring would alter the trajectory but not its general pattern.

The significance of this demonstration is that the state of the system at any particular future time looks simple to predict. All that is required is knowledge of the system's IC with perfect accuracy and infinite precision. As mentioned before, this is where the problems lie since this requirement is obviously not attainable. Even with instruments which could measure the block's initial velocity and potential energy extremely accurately, an infinitely precise measurement is impossible. Additionally, the

model of the system, a numerical solution to the differential equations, computes future states with only finite precision, which compounds error as the forecast evolves.

Suppose the dashed trajectory in Figure 1b represents the trajectory used to predict the true (solid line) evolution of the system. The IC used for prediction (open circle) are a little bit off from the true IC in both coordinate directions. This error is also a phase space vector which is the vector difference of the true and perturbed IC. The error vector points from the true IC to the perturbed IC.

The predicted trajectory then, begun from the slightly erred IC, never matches up to the true trajectory, so all predictions are in error. However, this system contains an *attractor* which is called a fixed point, causing all trajectories to converge (Tsonis and Elsner, 1989). This means that predictions, although always wrong, actually get closer to the truth further into the future. Unfortunately for weather forecasting, this is not the case for the atmospheric dynamical system.

An attractor is a region in a dynamical system's phase space that contains all states which naturally occur (Lorenz, 1993). If a system is forced to a state outside its attractor, the subsequent trajectory is drawn back into the attractor. Once in the attractor, the trajectory continues to evolve there unless the system is disturbed by an outside force. The most well known example of an attractor is the butterfly of the Lorenz system, a simple model of atmospheric convection (Figure 2). The region of this attractor consists of two distinct surfaces. A trajectory evolves within the attractor by spiraling outward on one surface then passing to the other surface, but at the same time never intersecting itself or another possible trajectory.

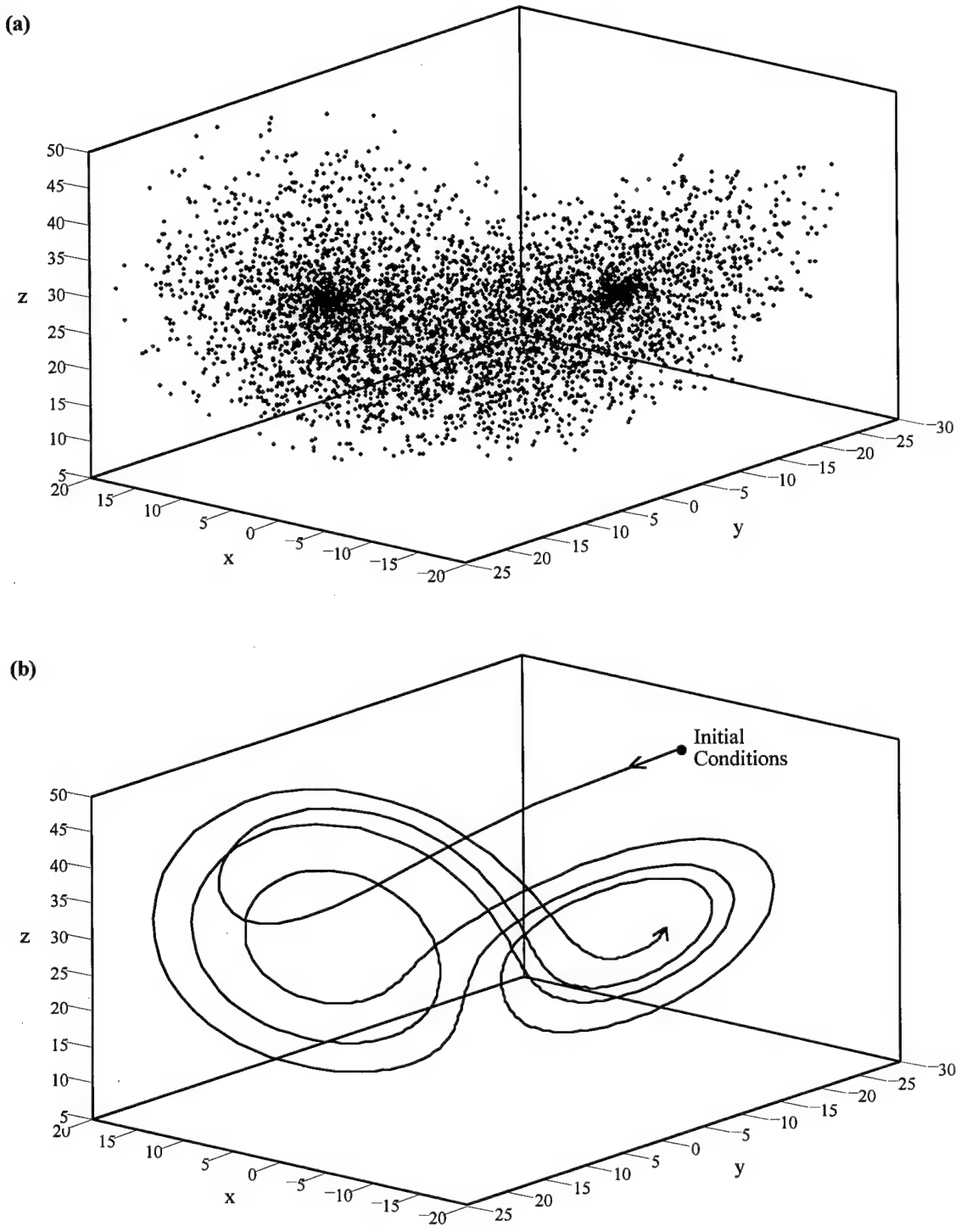


Figure 2. Phase space of the Lorenz system. (a) 5000 of the infinite number of naturally occurring states of the Lorenz butterfly (strange attractor of the dynamical system). (b) A trajectory started from outside of the attractor converges into the attractor and is then stuck evolving there.

Lorenz's butterfly is an example of a *strange attractor*. This term is rather misleading because such an attractor is common to the majority of chaotic dynamical systems, including the atmosphere (Lorenz, 1993; Tsonis and Elsner, 1989). Nearby trajectories evolving within a strange attractor usually diverge exponentially with time causing erred predictions to quickly get worse the further into the future a forecast is attempted. Dealing with this nonlinear error growth is the very essence of ensemble forecasting.

As will be described in the next section of this chapter, ensemble forecasting actually uses this problem (nonlinear error growth) to partially defeat the problem. Ensemble forecasting makes it possible for some of this nonlinear error growth to be accounted for, thus quantifying the uncertainty in a forecast (Toth and Kalnay, 1993). Since NCEP's MRF ensemble has demonstrated this ability, it was chosen as the foundation from which to produce calibrated PQPF for this thesis.

c. Ensemble Forecast*ing at NCEP*

The success of ensemble forecasting is highly dependent on two fundamental requirements concerning how the *n* perturbed IC are chosen (Toth and Kalnay, 1993). The requirements are based on the objective that the trajectory of each ensemble member must have an equally high probability of being the true trajectory of the atmosphere. Rephrased, the atmospheric trajectory should easily be a member of the ensemble. To accomplish this, each vector producing a perturbed set of IC must closely resemble the error vector of the analysis in both magnitude (requirement #1) and direction (requirement #2).

To demonstrate these requirements, Figure 3 gives a simplistic two-dimensional representation of atmospheric ensemble forecasting. What is the phase space dimension of the atmosphere? Equivalently, how big is the vector which completely describes a certain state of the atmosphere? In a typical primitive equation model, it takes eight variables to describe the atmosphere at a point. Each point, distributed horizontally and vertically around the earth, has a separate set of the eight variables. This puts the dimension of phase space on the order of 10^6 for a low-resolution global model. In theory, since the actual atmosphere has an infinite number of points, its phase space is of infinite dimension (Lorenz, 1993). The best that can be done for visualization of the atmosphere's phase space is a snap shot in time of a very limited subspace, which is commonly known as a weather chart. While displaying the basics of ensemble forecasting in only two dimensions as in Figure 3 may seem absurd, it is however meaningful because the concept of vectors and trajectories is the same no matter what the dimension.

The first requirement for successful ensemble forecasting is that the magnitude of the perturbations for the ensemble IC must be similar to the error magnitude of the analysis. While the actual error in any one analysis can never be known, the distribution of the error can be estimated from the climatic variance (Toth and Kalnay, 1993) or the difference between separate analysis/forecast systems (Hamill, 1998). By limiting the size of the ensemble perturbations to this distribution, all ensemble IC remain near both the analysis and the true IC of the atmosphere. The ensemble IC are like a cloud of plausible alternative analyses which encompass the best analysis (Wilks, 1995) and

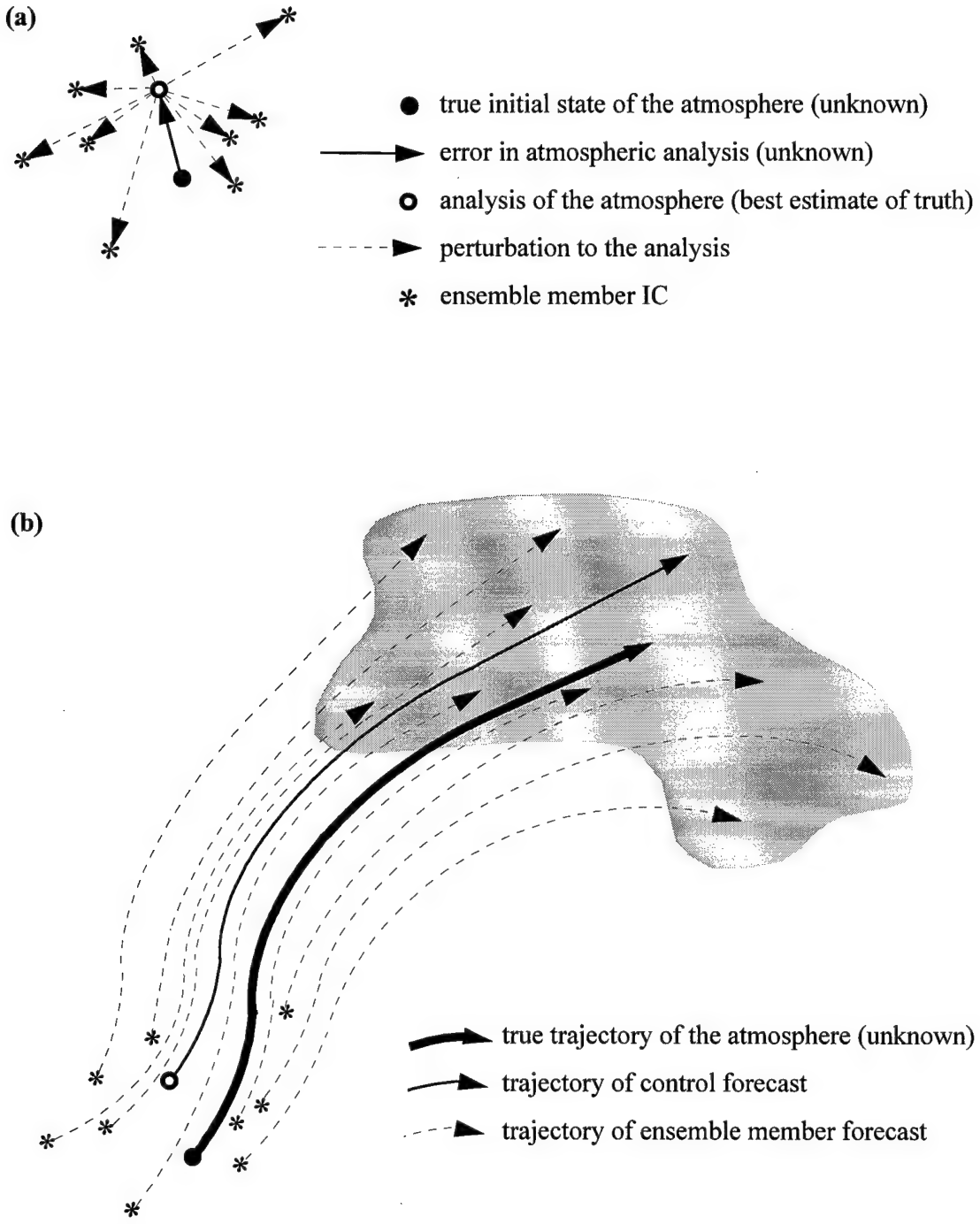


Figure 3. 2-D phase space schematic of an ensemble forecast. This ensemble consists of 10 members (the control forecast plus 9 perturbations). (a) Error in an atmospheric analysis is an unknown vector away from an unknown true initial state. Perturbations to the analysis which make up the ensemble IC are vectors away from the analysis with magnitude similar to the analysis error magnitude. (b) All trajectories are similar but diverge as time progresses. The true trajectory should lie within the shaded area, encompassed by the ensemble.

hopefully encompass the true IC (Figure 3a). There is no guarantee that the true IC will be encompassed since the actual vector of the analysis error is always unknown.

The second requirement for successful ensemble forecasting is that the perturbations must be distributed in phase space in such a way that the ensemble will likely encompasses the true trajectory as the forecast evolves (Figure 3b). In other words, the true trajectory must be a plausible member of the ensemble. If the true trajectory were to separate from the ensemble members, the ensemble forecast would be just as poor as the control forecast. The advantage of an ensemble forecast is that although the specific value of truth is not known, its area of possible values is known (shaded region of Figure 3b). With only a control forecast, which is known to be in error, it is impossible to know how far off or in which direction the truth may lie. The limit to the advantage of an ensemble forecast is that as the members continue to evolve, the ensemble variability eventually gets too high, making the truth's range of possibilities too great to be of value.

Meeting the first requirement for successful ensemble forecasting is fairly straightforward. The magnitude of the perturbations can easily be scaled to be within the estimated error distribution of the analysis. Conversely, there are many theories regarding the best method to meet the second requirement.

One way to make it likely that the ensemble forecast encompasses the truth is to simply start off with an extremely large number of random perturbations to the analysis with the correct magnitude. This generates a large number of ensemble members, increasing the chance of encompassing the truth. This method, called Monte Carlo, is however too impractical and inefficient because of the amount of computer power it

requires in order to get good results (Lorenz, 1993; Toth and Kalnay, 1993; Wilks, 1995). Only a method which uses a reasonable number of ensemble members can be applied operationally. This paper will now discuss the method used at NCEP in creating IC of the MRF ensemble which was used in this research.

Perturbations for the MRF ensemble are created by the method of Breeding of Growing Modes (BGM) (Toth and Kalnay, 1993). This method is designed to create highly variable ensemble trajectories by perturbing along *growing modes*. Given two states defined by phase space vectors $\vec{A}$ and $\vec{B}$, a mode is the vector difference in phase space between the states. For a growing mode, the trajectory from $\vec{A}$ must diverge from $\vec{B}$'s trajectory. A growing mode is really a two-way vector since if $\vec{B}$ is a growing mode away from $\vec{A}$, $\vec{A}$ must be a growing mode away from $\vec{B}$.

It was mentioned previously that within the atmosphere's strange attractor, trajectories usually diverge exponentially. This implies that almost any mode would be a growing mode. However, it is easy to define a perturbation which lies off the attractor but is of course still in the phase space. The trajectory of such a perturbation converges back to the attractor, thus making it a nongrowing mode. An example of this behavior is displayed back in Figure 2 on page 13. The IC in the figure would represent a nongrowing mode to any nearby state on or off of the attractor.

More importantly for ensemble forecasting, modes within the attractor diverge at different rates. For the highest ensemble variability, the mode with the maximum growth rate is desired. The BGM method estimates the maximum growth mode by a process similar to an analysis cycle (Toth and Kalnay, 1993).

The analysis cycle is the common method in NWP for producing the best possible analysis of the atmosphere for a single model run. In the period prior to a model's initial time, many short, sequential model runs are carried out to produce temporary first guesses to the true state of the atmosphere. At the end of each short model run, the first guess is nudged closer to truth by combining it with current observations. This process continues right up to the model's initial time when the final combination guess/observation state is used as the analysis. While this process does produce an analysis with an acceptably small error, the difference between the analysis and truth is believed to project increasingly onto a growing mode (Toth and Kalnay, 1993).

The analysis contains both random, nongrowing errors (convergent perturbations off the attractor) and organized, fast-growing errors which are growing modes (Toth and Kalnay, 1993). At the end of each short model run of the analysis cycle, the fast-growing errors dominate the total error of the first guess because the random errors decay or remain approximately the same size. When the first guess is combined with observational data, the error from growing modes is reduced but remains a major part of the analysis error. In the following short model run, these modes grow even further. In this way, a growing mode is bred so that by the end of the analysis cycle, the error in the analysis is a growing mode vector which extends from the true state of the atmosphere to the analysis, very close to the maximum growing mode.

To generate perturbations for an ensemble which resemble the errors in an analysis, the BGM method uses a breeding cycle which coexists with the analysis cycle (Toth and Kalnay, 1993). The process begins with an initial arbitrary perturbation to the

analysis (at first 00Z point in Figure 4). Next, the MRF model is run from both the analysis and the perturbed IC to produce two separate 6-hour forecasts. The difference (a phase space vector) in the two forecasts is then found. After the new analysis is produced, it is perturbed using a vector with the same direction as the forecast difference vector just calculated but scaled back to the size of the initial arbitrary perturbation. These steps are repeated so after a few short forecasts, the perturbation vector becomes a close estimate of the error in the analysis and thus also an approximation of the maximum growing mode.

There are several details of Figure 4 that must be noted. The two forecasts and the atmosphere's trajectory are always divergent. The exception to this is the initial perturbed forecast (first 00Z point) which originated from an arbitrary perturbation to the analysis. In this example, the initial perturbation is not along a growing mode as would normally be the case for a random perturbation. However, toward the end of the 6-hour forecast period, the fast-growing errors begin to show up, making the forecasts diverge. These fast-growing errors are then carried forward to the next perturbation so that after four cycles (second 00Z point of Figure 4), the vector for the perturbation is quite similar to the vector of the analysis error.

An ensemble composed of many of these growing mode perturbations is necessary to increase the probability of encompassing the true atmospheric trajectory. The NCEP MRF ensemble, containing a total of 17 members, is created by concurrently running seven independent breeding cycles (NCEP, 1995). Each cycle produces a similar but different estimate of the maximum growing mode. Recalling that a growing mode is

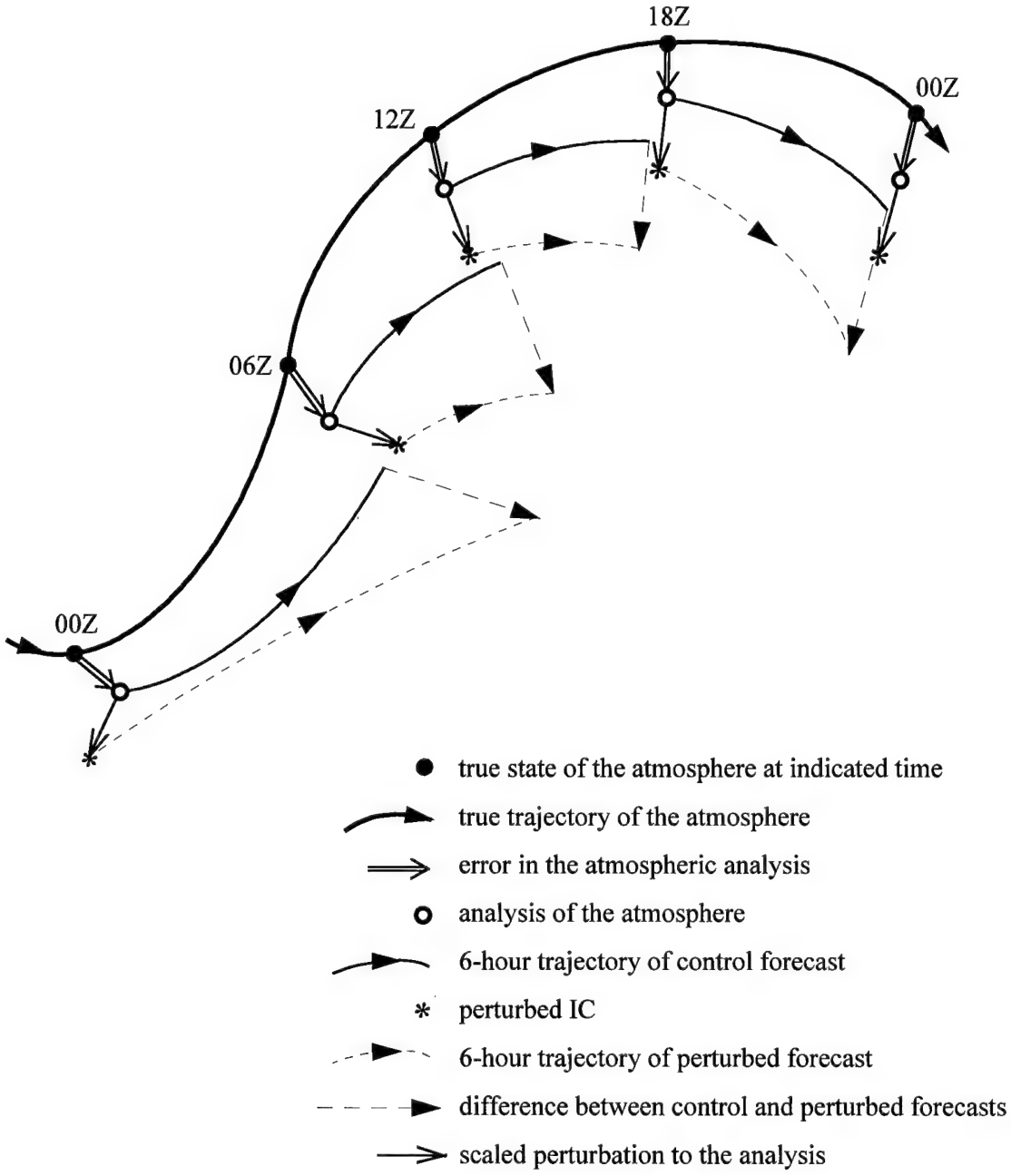


Figure 4. Schematic of the breeding cycle. The goal of this process is to get the perturbation (vector with single line) to be a maximum growing mode by forcing it to closely resemble the analysis error (vector with double line). The analysis error is close to the maximum growth mode by nature of the analysis cycle. Each new perturbation is determined by scaling down the vector difference between the control and perturbed forecasts at the end of a 6-hour forecast period.

a two way vector, one cycle actually produces two ensemble IC by adding and subtracting the mode to the analysis. So, the seven breeding cycles produce 14 of the 17 IC for the ensemble. The other 3 IC are control forecasts.

Due to limitations in computer resources, the ensemble initialized daily at 00Z actually consists of 12 members from 00Z model runs and 5 members run at the previous 12Z initial time. Of the 12 00Z members, 10 are bred perturbations and 2 are control forecasts run from the analysis at different model resolution. Of the 5 12Z members, 4 are bred perturbations and 1 is the control forecast from the analysis. The model runs are made out to 384 hours so with the 12Z time-lagged forecasts, each 00Z ensemble has a 372-hour or 15.5-day valid period.

In summary, NCEP's BGM method is an effective and efficient way to generate an ensemble which meets the fundamental requirements of successful ensemble forecasting (Toth and Kalnay, 1993). The perturbations of the MRF ensemble IC truly represent the errors that likely exist in the analysis. The ensemble trajectories all diverge at near maximal rate making it probable that the true trajectory of the atmosphere is encompassed.

3. *Experimental Methodology*

a. Overview

This chapter details the methodology involved in this research. It begins with information on the data that were used, how the data were processed, and possible limitations of the data. Next, the chapter shows the theory behind Anderson's (1996) binned probability ensemble (BPE) technique and its application in examining the MRF ensemble precipitation data for systematic errors. Lastly, different methods for producing PQPF are described with a focus on the theory for the calibrated PQPF which follows the work of Hamill and Colucci (1997).

b. Research Data

1) Ensemble Data

Probabilistic forecasts for any weather parameter can be produced from ensemble model output. This research was applied only to precipitation over the conterminous United States for reasons of observational data availability and for comparison to a different research project carried out at NCEP. NCEP concurrently worked on a project with the same goal as this research but with an entirely different approach (Toth et al., 1998). NCEP's approach is briefly described in section d of this chapter. By using the same weather parameter (precipitation) and geographical location (US), the results of the two projects could be directly compared.

To construct the calibration for this Thesis and test its reliability, an archive of many ensemble forecast cases was needed. A total of 358 daily MRF ensemble forecast

cases from the period SEP 96 through NOV 98 were used. Many forecast case days over this period were missing. The available data was downloaded in gridded binary (grib) format from an archive at the Climate Diagnostics Center, an office of the National Oceanic and Atmospheric Association. With each case day giving 17 separate global precipitation forecasts, each valid every 12 hours over a 16-day valid period, the information totaled approximately 4 gigabytes in size.

Fortran programs were used to process this large amount of data into a useful format. First, programs provided by NCEP were used to decode the grib data. Next, another program reconstructed the ensemble. Recall that each 00Z MRF ensemble consists of 5 members from the previous 12Z model run and 12 members from the 00Z model run. Since the downloaded data gave the 17 original forecasts, valid times of the two sets of forecasts had to be correctly matched up to reconstruct the ensemble.

Concurrently with this reconstruction, the forecast data over the 25x11 MRF 2.5° subgrid over the US used in this research (Figure 5a) was separated out from the global data. Latitudinal spacing between grid points is 278 km. Longitudinal spacing between grid points is 252 km at row J = 1 and decreases by the cosine of the latitude to 179 km at row J = 11.

Further processing was required to get the forecast data to match up with the observational data. The observational data used, described in the next section, gives 24-hour cumulative precipitation (*pcp24*) valid every 12Z instead of 12-hour cumulative precipitation which is forecast by the MRF ensemble. For this reason, the 12Z – 00Z and 00Z – 12Z forecasts for every day of the forecast valid period were added together to

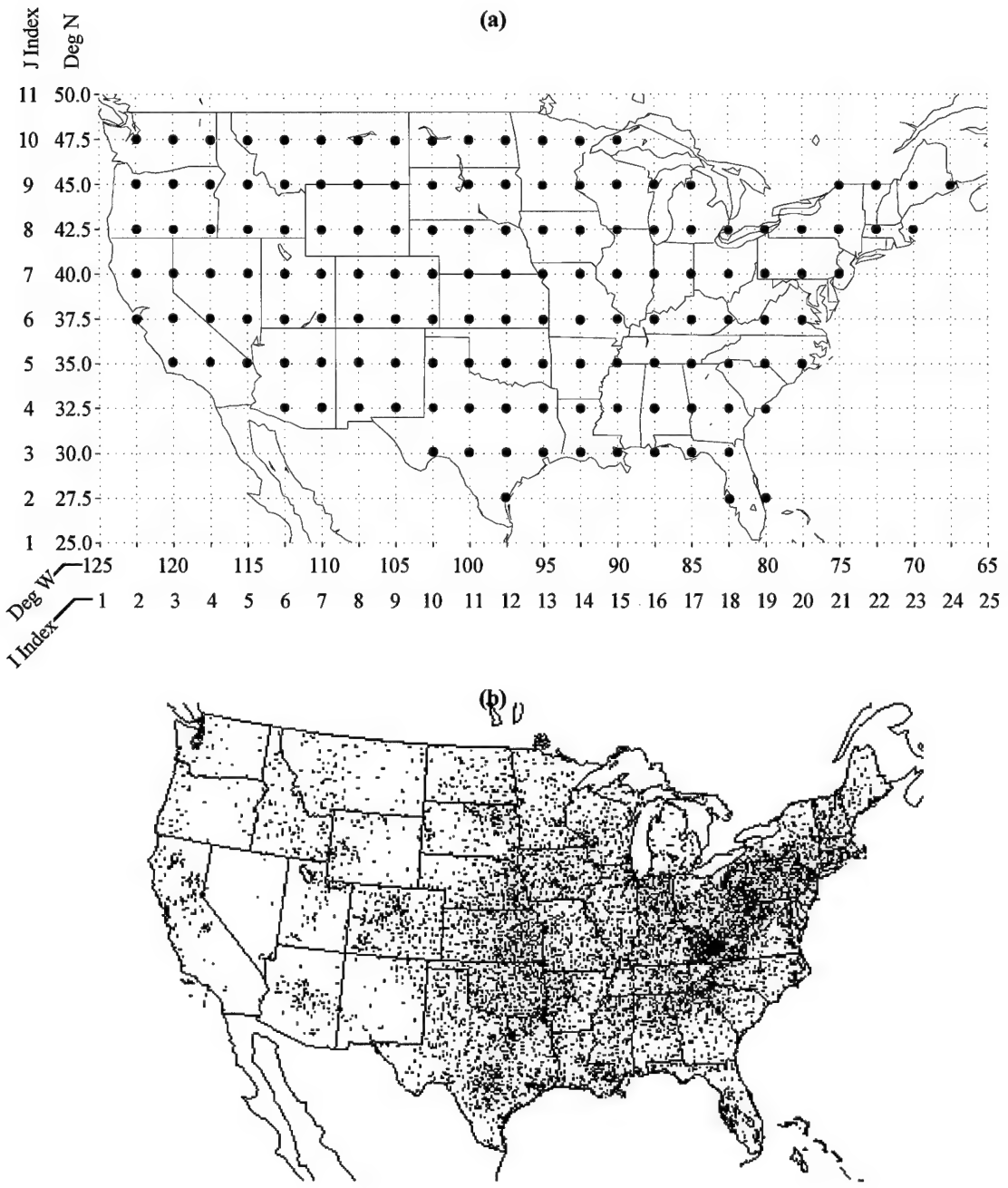


Figure 5. Forecast region of the research. (a) 25x11 MRF 2.5° subgrid for ensemble forecast data. Grid points with dots are locations where precipitation observations were available for verification of the ensemble forecasts. (b) Network of rain gages belonging to the National Weather Service's River Forecast Centers.

make a *pcp24* forecast at each 12Z point. Thus each ensemble member consists of 15 *pcp24* forecasts valid at the 36-hour, 60-hour, …, and 372-hour valid times. Note that the first 12Z valid point of the ensemble can not be used since with a model initial time of 00Z, there is only one 12-hour cumulative precipitation forecast at the 12-hour valid point. The 15 valid points of the ensemble are hereafter referred to in days (E.g.: The first valid point at 36 hours, is a 1.5-day forecast or a forecast with a 1.5-day lead time).

The full data set of 358 forecast case days was divided by month into a training data set and a forecasting data set (Table 1). The training data set was used exclusively for construction of the calibration. The forecasting data set was used in generating PQPF after development of the calibration, so forecasts were made with no prior knowledge of the verification value. This allowed for fair evaluation and comparison of PQPF quality.

2) Observational Data

In developing a calibration for PQPF, an important aspect was the choice of verification data. At first, it seems peculiar to have to choose between different values of truth. This dilemma, presented earlier in this thesis in a different context, comes about because the truth can never really be known. It follows then that there may be more than one estimate of what is considered truth. Since the goal here was to account for systematic errors, different truths would likely indicate different errors and result in different calibrations.

The choice of truth was particularly difficult for this research because of the weather parameter being forecast, precipitation. Suppose that a *pcp24* forecast for Vandenberg AFB, CA, was 5.5 mm. The most obvious way to verify such a forecast is

Table 1. Division of forecast case days into training and forecasting data sets.

Month	Training Data	Forecasting Data	Dates of Ensemble Forecast Case Days
SEP 96		X	20 21 22 23 24 25 26 27 28 29 30
OCT 96	X		2 3 4 5 6 7 8 9 10 11 12 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
NOV 96		X	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26 27 28 29 30
DEC 96	X		1 2 3 4 5 6 7 8 9 10 12 13 14 15 16 17 18 19 21 22 23 24 25 26 27 28 29 30 31
JAN 97		X	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 19 20 21 22 25 26 27 28 30 31
FEB 97	X		1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 28
MAR 97		X	1 2 3 4 5 7 10 12 13 14 15 16 17 18 19 20 21 23 24 25 26 28 29 30 31
APR 97	X		1 3 5 6 7 8 9 10 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30
MAY 97		X	1 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26 27 28 29 30
JUN 97	X		1 2 3 4 5 6 7 8 9 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
JUL 97		X	1 2 3 4 5 6 7 8 9 10 11 13 14 15 16 18 19 21 22 23 24 25 26 27 28 29 30
AUG 97	X		1 2 3 4 5 6 7 8 11 12 13 14 15 16 17 19 20 21 22 23 24 25 27 28 29 30 31
SEP 97		X	1 2 6 7 8 9 10 11 12 13 14 15 18 23 27 28 29 30
OCT 97	X		3 4 5 6 12 13 14 16 19 20 24 25 26 27 28
NOV 97		X	4 6 7 8 9 10 13 14 15 19 20 21 22 23 26

to use the air terminal's rain gage total over the forecast period. However, because of the high spatial variability of rainfall, the gage's total could easily be significantly different than the forecast value even though the forecast was excellent (Baldwin, 1997). What is needed is a precipitation total which is consistently representative for the local area about the point of interest. This can be provided by averaging the totals from many nearby gages.

Since this research worked with output from the MRF, representative precipitation totals at the MRF grid points were desired. NCEP provided *pcp24* observations valid daily at 12Z for MRF 2.5° grid points within the conterminous United States for SEP 96 through NOV 98. This data was prepared by NCEP using reports from a network of approximately 10000 rain gages belonging to the National Weather Service's River Forecast Centers (Figure 5b) which were first spatially averaged onto a 40 km grid. The 40 km grid was then remapped by NCEP to the MRF 2.5° grid.

NCEP's remapping process is schematically depicted in Figure 6 (Baldwin, 1997). The first step is to divide up each original grid box into 16 sub grid boxes which each take on the original box's *pcp24* value. Next, the new grid is overlaid onto the original grid. The *pcp24* value for a new grid box is calculated by taking an area weighted average of all the values from sub grid boxes whose centers fall within the new grid box.

The uncertainty in rain gage measurements is ± 0.01", or ± 0.25 mm (ABRFC, no date). In the spatial averaging step and the subsequent remapping, the value of *pcp24* in the hundredths digit is carried along in the computations. The final *pcp24* should therefore be rounded off to the nearest 0.25 mm to reflect the true precision of the data.

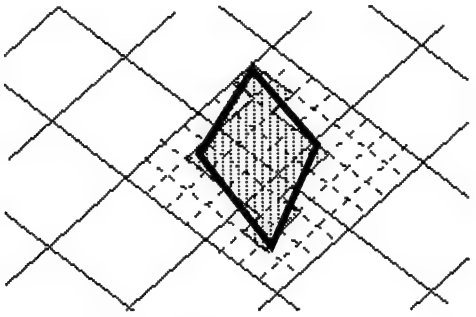


Figure 6. (from Figure 4 in Baldwin, 1997) Schematic of the remapping process. Thin solid lines show grid boxes from the original spatial averaging of rain gage data. Dashed lines break up each original grid box into 4x4 sub grid boxes. Thick solid lines outline a box of the new grid being remapped to. Shaded sub grid boxes are used to get a value for the new grid box.

However, since the ensemble forecasts are made to the nearest 0.1 mm, it was decided to preserve the unrealistic 0.01 mm precision of observed pcp_{24} to reduce the number of ties between forecasts and verification. This has a positive impact on the PQPF calibration which will become evident later in this chapter.

Another alternative choice of truth to be used for verification purposes was seriously considered. NCEP has developed a multisensor analysis of cumulative precipitation which combines data from the rain gage network described above with data from the nationwide network of NEXRAD sites. The resulting cumulative precipitation analysis should in theory be more accurate because of the extremely high resolution of information. However, a major difficulty lies in estimating precipitation rates from the radar returns. Since the reliability of the multisensor analysis is as yet unproven, it was not chosen to represent truth.

A possible source of error for this research came from using the spatially averaged rain gage data for verification purposes. The error stems from the fact that the theoretical

distribution of precipitation typically follows a gamma distribution bounded on the left by zero (Wilks, 1995). For wet events of widespread precipitation, the gamma distribution has a large value of the gamma shape parameter, α. Since such a distribution is near normal, the spatial average is valid as a representative value with the mean being the precipitation value most likely to occur. Conversely, events of very low precipitation have a gamma distribution with a small α making the distribution more exponential. The mean would then be too high to be a representative value for the event since the most likely value of precipitation would be closer to zero.

c. Systematic Error

The need for this research was based on the premise that the MRF ensemble contained significant systematic errors. Once these errors were found, some sort of calibration could then be developed. If the systematic errors were insignificant or not present at all (e.g., a perfect model and an ideal choice of ensemble perturbations), the ensemble would be described as *well calibrated.* The resulting PQPF of such an ensemble would display perfect reliability, and there would be no need for further calibration. For this reason, the first step of this research was to investigate the MRF ensemble precipitation forecasts for systematic errors.

To understand what is meant by systematic error from the point of view of chaos theory, refer back to Figure 3b on page 16. For a well calibrated ensemble, the true state of the atmosphere will most likely lie in the shaded region, encompassed by the ensemble members. Using an ensemble with an imperfect model and less than ideal perturbations

causes problems. The ensemble trajectories would still all behave in a similar manner, but this behavior would be slightly different compared to the atmosphere's trajectory. The result, shown in Figure 7, is that the region in phase space encompassed by the ensemble members is shifted with respect to the truth. The ensemble may still encompass the true state of the atmosphere, but with less likelihood.

Further comparison of Figure 3b and Figure 7 reveals another important distinction. Recall that trajectories evolving chaotically can not intersect themselves or any other trajectory, thus avoiding periodic motion. Note in Figure 7 that while the

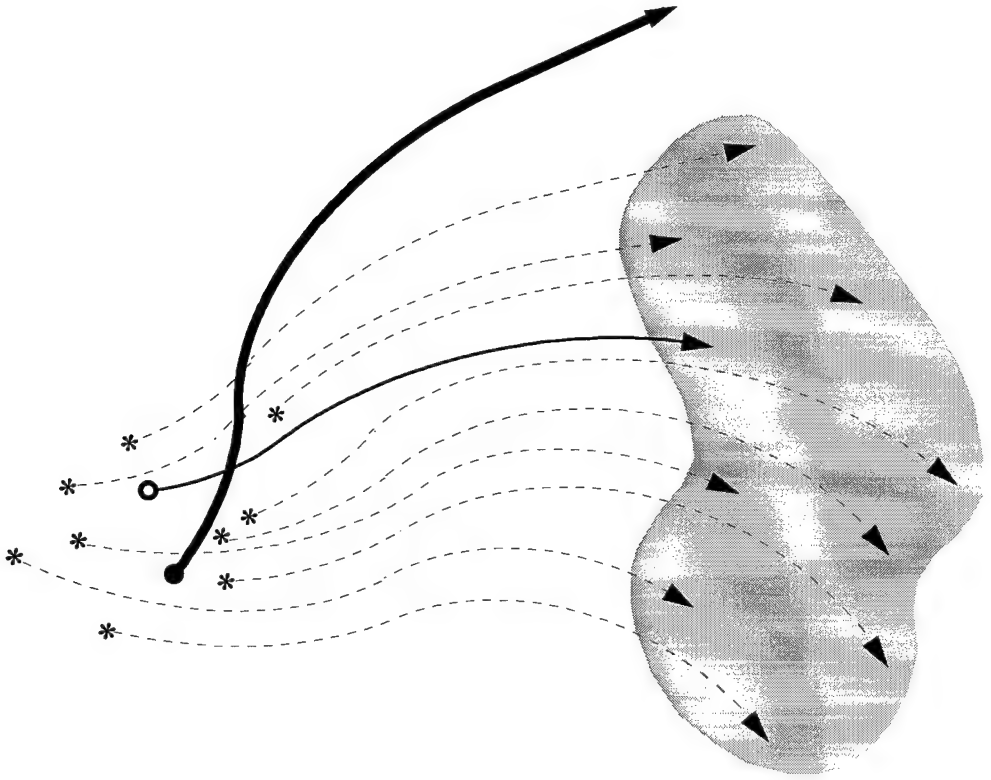


Figure 7. Schematic of a forecast from the same IC as in Figure 3 using a poorly calibrated ensemble. The trajectory of the atmosphere is necessarily exactly the same. Trajectories of the ensemble members behave quite differently than the atmosphere because of both an imperfect model and poorly chosen ensemble perturbations. The ensemble then fails to encompass the truth.

ensemble members' trajectories do not intersect themselves, several do intersect the atmospheric trajectory. This happens because the ensemble trajectories are governed by a set of rules (the model) different from those of the atmosphere. The dynamical system of the model is an erred approximation of the atmospheric dynamical system, precisely the reason why the truth is less likely to be encompassed by the ensemble. The better the model and choice of ensemble perturbations, the more Figure 7 looks like Figure 3b.

The true goal of this research can now be clearly seen. Rather than try to reduce the error by making a better model and/or improving the ensemble perturbation scheme, this research demonstrates improvements achieved by correcting systematic error. The investigation of such error was actually an attempt to define the shifting of the ensemble region away from truth. Once this shift was defined, it was used to interpret the ensemble data in order to make a calibrated PQPF. This process was complicated by the fact that the shift varies depending upon the region of phase space (i.e., the atmospheric conditions). For this procedure to be useful, some stationarity in the shift had to be found. If the shift were totally random, it would be impossible to design a calibration. Thus it was necessary to identify conditions that repeatedly led to similar shifts. This is thoroughly discussed in the next chapter.

The main tool for investigating the systematic error was the verification rank histogram created with the BPE technique. There are several terms involved here which need to be defined. The *verification* is the observed value of the forecast variable. A *rank* is simply the ordinal place number of a value among other values which are arranged from smallest to largest. A *bin* is a possible rank of the verification when it is pooled and

ranked among the ensemble forecast values. Therefore, an ensemble of n members contains $n + 1$ bins. A bin covers the range of forecast variable values which exist between the values of two ranked members of an ensemble, or beyond either extreme value of the ranked members.

Figure 8 shows an example of 5 bins and their range of verification values for a hypothetical 4-member ensemble of 12-hour precipitation forecasts. Recall that each member of the ensemble is forecasting for the same event. If this forecast verified with an observed 12-hour precipitation of 2.1 mm, the verification rank would be 2.

In the event that the verification exactly equals one or more of the ensemble members, the rank is randomly assigned among its possible values (Hamill and Colucci, 1997). Suppose the verification value for the forecast in Figure 8 was 5.8 mm. The verification rank would then randomly take on one of its two possible values, a 3 or a 4.

With a perfect model and an ideal set of ensemble IC, the verification has an equally likely chance of falling into any of the $n + 1$ bins, giving verification ranks a uniform distribution (Anderson, 1996). This is counterintuitive since the verification

Ranked Ensemble Members	{ 0.9 ,	3.2 ,	5.8,	9.2 }	
Bin #	1	2	3	4	5
Verification Range	$0.0 \le V < 0.9$	$0.9 \le V < 3.2$	$3.2 \le V < 5.8$	$5.8 \le V < 9.2$	$9.2 \le V$

Figure 8. Example of verification bins of the BPE technique for a 4-member ensemble of 12-hour precipitation forecasts in mm. The value of each ensemble member represents a break in the verification range of the bins. Notice that the ranges are unequal in size.

range of each bin is not the same size. The reason for the uniform distribution is that while the probability distribution of the variable may not be uniform, the process which generates values of the random variable (ensemble or observation values) is uniform. Values of the variable are tied to quantiles through the variable's cumulative distribution function (CDF). It is the quantiles which vary uniformly.

This key point of the BPE technique is illustrated in Figure 9 for a hypothetical gamma distribution of 12-hour cumulative precipitation (the random variable) at some location. Samples of the random variable are generated through random quantiles which have equal (uniform) chance of being any value between 0 and 1. A quantile corresponds to a value of the random variable on the CDF curve. (E.g.: In Figure 9a, the 0.85 quantile gives a cumulative precipitation of 4.7 mm.) Bins for the verification, previously described as intervals of the random variable, are actually quantile intervals of equal size when averaged over many samples. Given any two quantile intervals of equal size, a single sample has an equally likely chance to occur in either interval (bin). Because the CDF is nonuniform, the equal-size quantile intervals do not correspond to equally sized intervals of the random variable (Figure 9a).

With a large set of independent samples of ensemble forecasts and verifications, a verification rank histogram can be built. It shows the number of times the verification occurred in each rank (bin) over the entire sample space. The histogram can be tested for uniformity using the χ^2 goodness-of-fit test with p value significance of 0.01. A histogram failing this test is an indication that systematic errors exist in the ensemble data (Anderson, 1996). As will be described in the next section of this chapter, the

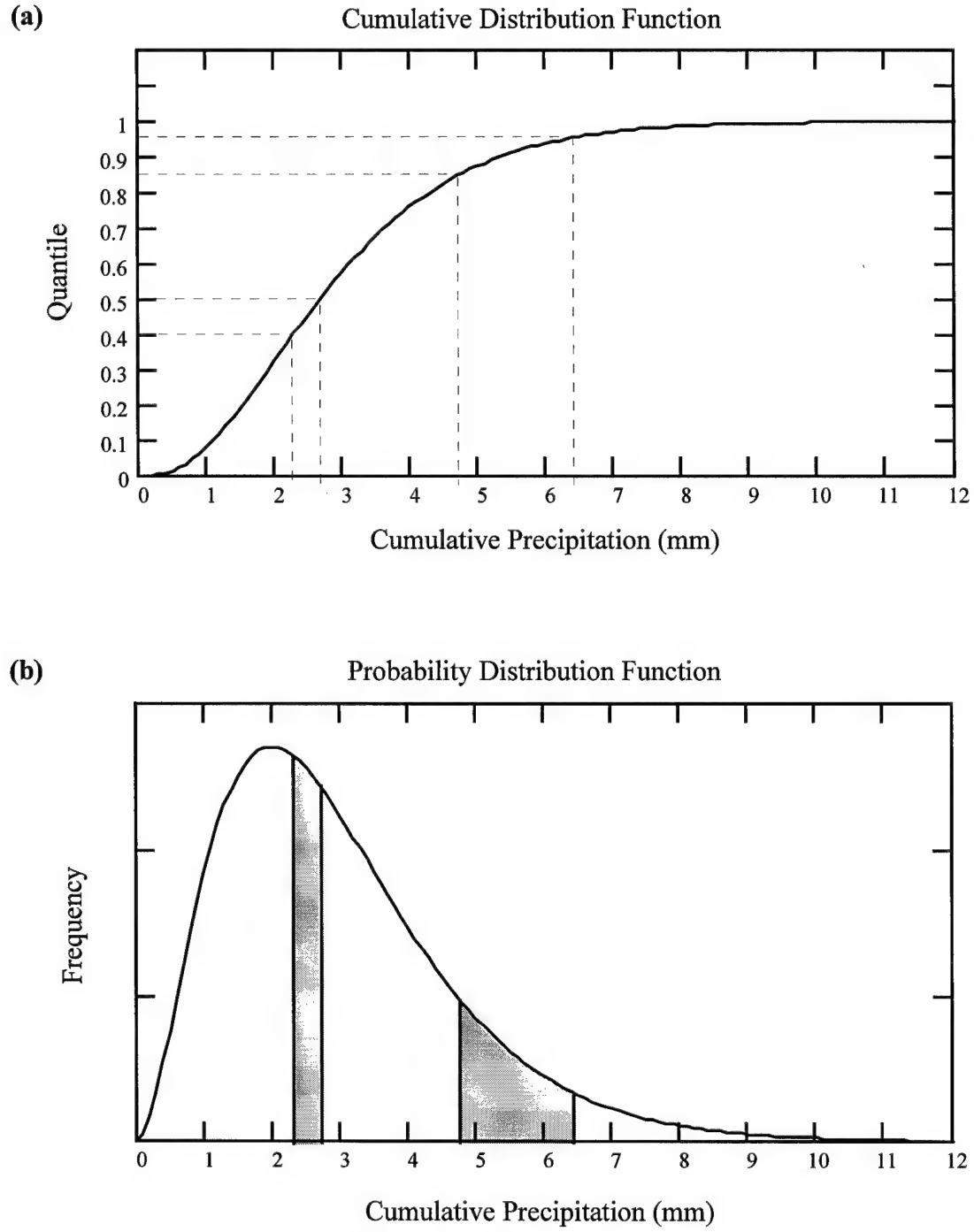


Figure 9. Demonstration of equal bin probability of the BPE technique for a perfect model and an ideal set of ensemble IC, for a gamma distribution of 12-hour cumulative precipitation. (a) Two possible bins are represented by quantile intervals of equal size (0.1) on the CDF graph. These correspond to unequal cumulative precipitation ranges. Probability that a value of cumulative precipitation will occur in either range is 1/10. (b) The two ranges also bound equal areas under the probability distribution curve. Each shaded area represents a probability of 1/10.

nonuniform verification rank can then be used to produce calibrated probability forecasts (Hamill and Colucci, 1997).

An example of such a histogram from this research is displayed in Figure 10 for the results of verification ranking for ensemble forecasts with 2.5-day lead times. Clearly, the occurrences over the ranks are far from uniform showing that the ensemble has significant systematic errors. Since ranks #1 and #18 got by far the highest number of occurrences, the ensemble often completely underforecast or completely overforecast the verification value of precipitation. Histograms of this sort are analyzed in more detail in chapter 4.

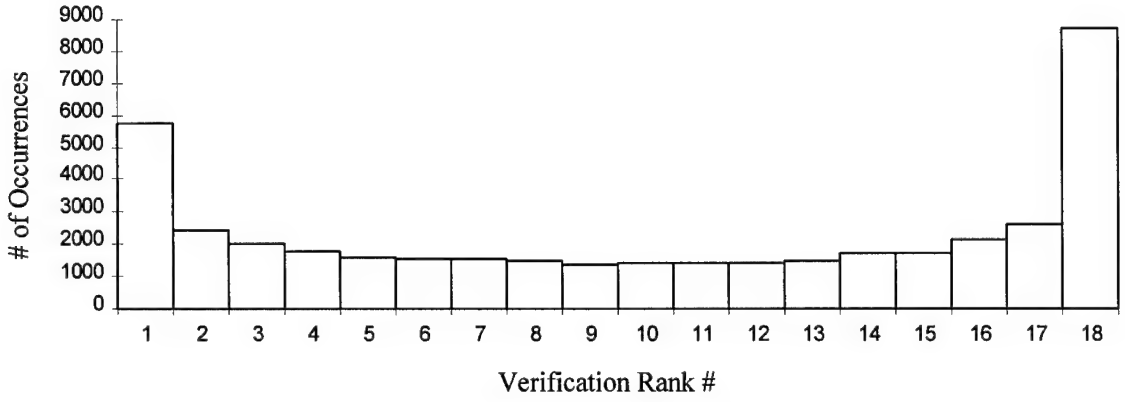


Figure 10. Verification rank histogram for 2.5-day *pcp24* forecasts from a sample of 42065 ensemble forecasts.

d. Producing PQPF from the Ensemble

Even with no knowledge of the nonuniformity of verification rank (i.e., presence of systematic errors), skillful but poorly calibrated probability forecasts can still be derived from an ensemble with systematic errors. As a basis for comparison, two such methods were applied to produce PQPF from the MRF ensemble. These first two methods, which should be considered approximations, will be called the *democratic voting* method (Doran, 1997) and the *uniform ranks* method.

NCEP developed a methodology designed to account for the systematic errors (Toth et al., 1998). NCEP's approach to calibrating PQPF, which has not yet been fully implemented, is to adjust the distribution of the ensemble precipitation forecasts to more closely match the observed distribution of precipitation. The process involved is to fit both the ensemble and the observations to a three-parameter gamma distribution:

$$f(x;\alpha,\beta,\gamma) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)}(x-\gamma)^{\alpha-1}\exp\left(\frac{-(x-\gamma)}{\beta}\right) \qquad (1)$$

for each grid box over the US. That is, each grid box produces a set of ensemble vs. observed parameter values (α, β, γ) for two different gamma distributions. For an ensemble containing systematic errors, the difference between the two distributions is similar over the entire sample space. Assuming that the stochastic process producing the systematic errors is stationary, the averaged difference in the two sets of parameters can be used to correct the distribution of future ensemble forecasts and arrive at PQPF with improved calibration.

This research applied a completely different method from NCEP's approach to account for systematic errors. This method, termed the *weighted ranks* method in this thesis, was introduced by Hamill and Colucci (1997). Its basic idea is to use the information in the nonuniform verification rank distribution(s) to produce calibrated probabilistic forecasts which account for systematic errors.

Before obtaining any probabilities, categories for the forecast variable must be defined. The ranges for pcp_{24} categories used for this research are identical to the ones in operational use at NCEP in reporting PQPF. The categories are CAT1: $pcp_{24} \geq 0.1$", CAT2: $pcp_{24} \geq 0.25$", CAT3: $pcp_{24} \geq 0.5$", and CAT4: $pcp_{24} \geq 1.0$". Since this research worked with SI units, the category thresholds are 2.54 mm, 6.35 mm, 12.70 mm, and 25.40 mm respectively. Notice that the categories are designed to give *decumulative* probability rather than mutually exclusive and collectively exhaustive (MECE) probability.

With decumulative type categories, the ranges of the forecast variable in each category overlap. For one particular forecast, probabilities for all categories do not sum to 1.0 and must decrease or remain the same as category threshold increases. It should also be noted that the choice of the number of categories, their thresholds, and their type (MECE vs. decumulative) has no impact on the construction of the calibration. The somewhat arbitrary categories are simply a useful way to present probabilistic forecasts.

For demonstration of differences between category probabilities produced by the three methods, consider an actual example of a MRF ensemble pcp_{24} forecast (in mm):

$$ENS = \{0.8, 1.3, 6.8, 7.1, 7.7, 8.7, 9.0, 10.2, 11.0,$$
$$11.2, 14.9, 14.9, 16.2, 19.2, 20.6, 23.0, 24.0\}$$

where each of the 17 members is an alternative forecast of the same event. In this case, the event was *pcp24* from 12Z on 13 OCT 97 to 12Z on 14 OCT 97 at grid point (16,6) in western Kentucky (see Figure 5, page 25). The model initial time was 00Z on 12 OCT 97 so the forecast period corresponds to the ensemble forecast lead time of 2.5 days.

Obtaining probability in each category ($P_{CAT\#}$) using the democratic voting method is the most straightforward method. As the name implies, each ensemble member gets an equal vote on which categories the verification will occur in. Mathematically, the number of ensemble members which fall into a category's range are first tallied. Dividing this result by the total number of members yields the probability for that category. Table 2 shows the results of this process for *ENS*. The vote from members that forecast below the lowest category's threshold is ignored in practice but included here for completeness as P_{NoCAT}. Because this is the method currently used at NCEP to derive PQPF, reliability of its PQPF was considered the benchmark from which to measure the improvement of the calibrated PQPF produced from the weighted ranks method.

Table 2. Computation of category probabilities by the democratic voting method for the example MRF ensemble *pcp24* forecast, *ENS*.

Category	*pcp24* Range (mm)	Number of Ensemble Members in Range	Probability Calculation
none	< 2.54	2	$P_{NoCAT} = 2/17 = 0.12$
CAT1	≥ 2.54	15	$P_{CAT1} = 15/17 = 0.88$
CAT2	≥ 6.35	15	$P_{CAT2} = 15/17 = 0.88$
CAT3	≥ 12.70	7	$P_{CAT3} = 7/17 = 0.41$
CAT4	≥ 25.40	0	$P_{CAT4} = 0/17 = 0.0$

A slightly better but more involved approximation of probabilities is obtained using the uniform ranks method (Hamill and Colucci, 1997). It was included in this research as a simplification of the weighted ranks method and another benchmark for comparison. The method begins by assuming that the ensemble contains no systematic errors. Probability is assigned to the four categories by summing the probabilities from appropriate ranks of a verification rank histogram with uniform probability in each rank (Figure 11a). Which ranks' probabilities to sum for a particular category is based on where the category threshold falls among the ranked ensemble members. As with the democratic voting method, some probability may not fall into any category (P_{NoCAT}).

To better understand this process, consider category 3 probability in detail. The question is: What is the probability that the verification will exceed the category threshold of 12.7 mm? In Figure 11a, the dashed lines between the ranked ensemble members are the verification ranks (i.e., possible ranked placements of the verification within the ensemble). The category 3 threshold is in rank #11, so if the verification were to occur in any rank greater than 11, the threshold would be exceeded. Therefore, the probabilities of the verification occurring in each of ranks #12 – #18 are summed. For the uniform ranks method, the probability in each rank is simply 1/18 since it was assumed that the verification has an equally likely chance of falling anywhere among the 17 ordered ensemble members. This makes $P_{CAT3} = 7 \cdot (1/18) = 0.39$, the bulk of the probability for category 3. A bit more probability comes from the fact that the verification can exceed the threshold if the verification occurs in rank #11. This additional probability is what makes this method a better approximation than the democratic voting method.

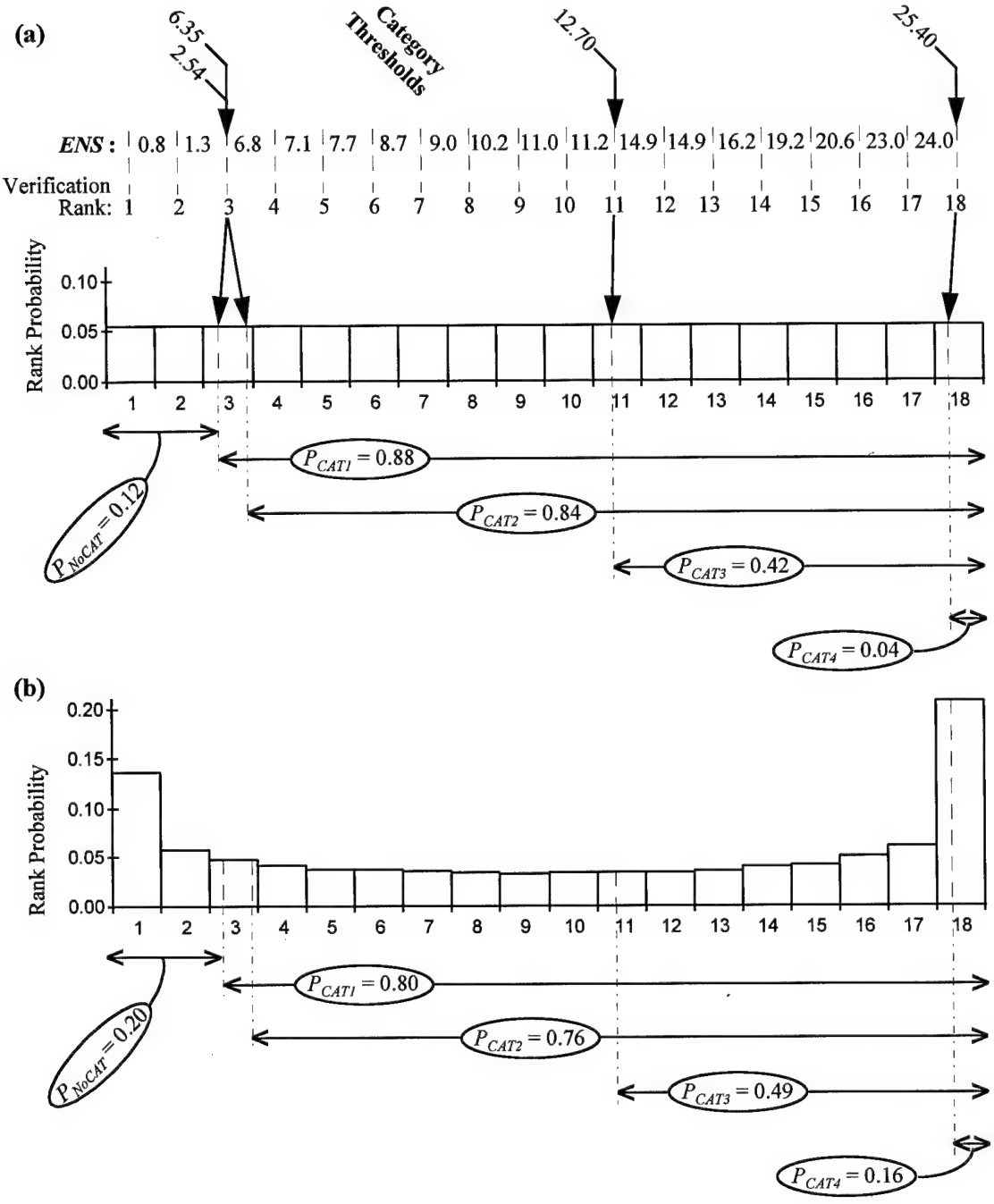


Figure 11. Calculation of PQPF for each *pcp24* category from the sample ensemble forecast, *ENS,* using the (a) uniform ranks method, and (b) weighted ranks method. The position of a category's threshold among the ranked ensemble members determines the starting point. The category probability is found by summing each rank's probabilities in the histogram to the right of the starting point. The starting points for each category are identical for both methods. However, the two methods produce different results since the probabilities of the ranks which get summed for (a) are much different than for (b).

When a category's threshold falls between two members, the portion of probability from the corresponding verification rank is broken out by Equation 2 (adapted from Equation 4, Hamill and Colucci, 1997) to be summed with the probabilities from the ranks described above.

$$P\left(T < V < x_{i+1}\right) = \left(\frac{x_{i+1} - T}{x_{i+1} - x_i}\right) \cdot RP_{i+1} \tag{2}$$

where T is the threshold value, V is the verification value, x_{i+1} is the value of the ensemble member with rank $i + 1$, x_i is the value of the ensemble member with rank i, and RP_{i+1} is the amount of probability in verification rank $i + 1$. This step assumes that the random variable *pcp24* is uniformly distributed between ensemble members. (Note: Equations adapted from Hamill and Colucci (1997) were altered to give decumulative probability.)

Returning to the category 3 example, approximately 6/10 of the probability in verification rank #11 is added to P_{CAT3} since the category 3 threshold of 12.70 is six tenths of the way from 14.9 toward 11.2. So for the total probability, $P_{CAT3} = 0.39 + (6/10) \cdot (1/18) = 0.42$. According to the uniform ranks method there is a 42% chance of *pcp24* >12.7 mm.

When a category's threshold is in the highest verification rank, as for category 4 in Figure 11, a much different procedure is followed. The probability is found by taking a portion of probability in rank #18 based on the numerical distance between the highest member and the category threshold. The probability of rank #18 is considered to be the upper extreme end of the sample's theoretical Gumbel CDF: (from Equation 4.43, Wilks, 1995).

$$F(x) = \exp\left[-\exp\left(\frac{\xi - x}{\beta}\right)\right] \tag{3}$$

where x is the random variable, and β and ξ are the Gumbel parameters estimated with the equations: (from 4.44a and 4.44b, Wilks, 1995)

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi} \tag{4}$$

$$\hat{\xi} = \bar{x} - \gamma\hat{\beta} \tag{5}$$

where s is the sample standard deviation, $\bar{x}$ is the sample mean, and γ is Euler's constant. The Gumbel distribution was used because of its ability to characterize extreme events (Hamill and Colucci, 1997; Wilks, 1995). The probability value for the category is found with Equation 6 (adapted from Equation 5, Hamill and Colucci, 1997).

$$P(T < V) = \left(\frac{1 - F(T)}{1 - F(x_{17})}\right) \cdot RP_{18} \tag{6}$$

where V is again the verification value, $F(T)$ is the Gumbel CDF value at the threshold value T, $F(x_{17})$ is the Gumbel CDF value at the value of the highest ranked ensemble member, x_{17}, and RP_{18} is the amount of probability in verification rank #18.

In Figure 11a, probability of 0.04 goes into P_{CAT4} from verification rank #18. Figure 12 shows a graphic display of this computation. This probability reflects the chance that the verification may occur in category 4 even though all the ensemble members predict values below the category 4 threshold. For the opposite extreme, a threshold falling below the lowest ensemble member, the portion of rank #1 is found using Equation 2 where 0.0 is used for the value for x_i. This technique assumes a uniform distribution between 0.0 and the lowest ensemble member (Hamill and Colucci, 1997).

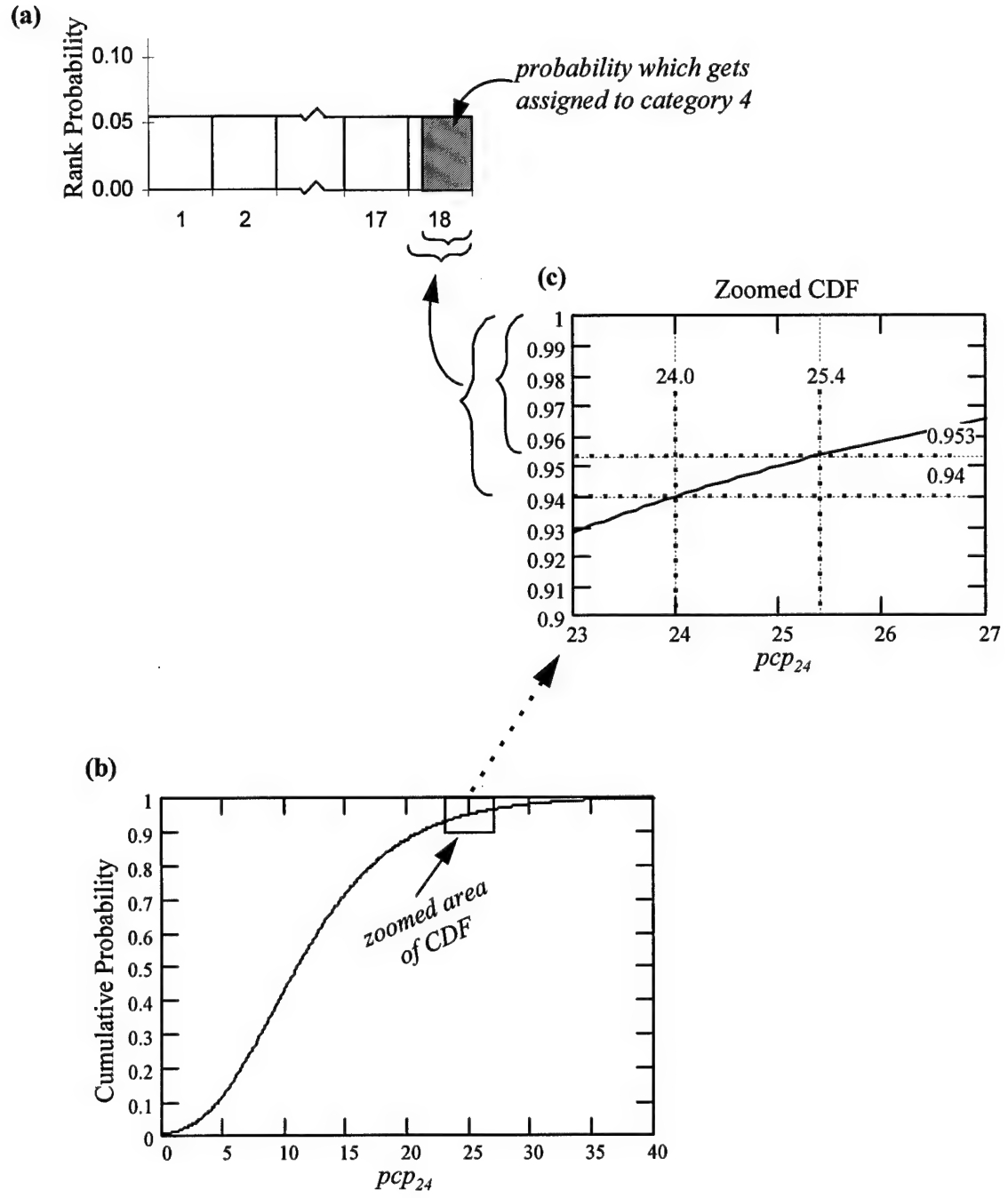


Figure 12. Determination of probability for a category whose threshold exceeds the highest ensemble member. (a) Verification rank probabilities for uniform ranks method. The fraction of rank #18 probability for category 4 is determined on the CDF. (b) Full view of the fitted Gumbel CDF for the sample ensemble forecast, *ENS*. (c) Zoomed view of CDF with markers for highest forecast pcp_{24} value (24.0) and category 4 threshold (25.4), and corresponding quantiles, 0.94 and 0.953. The ratio $(1 - 0.953) / (1 - 0.94) \approx 0.8$, so 4/5 of the probability of rank #18 is the PQPF for category 4.

The weighted ranks method should theoretically produce the most reliable PQPF (Figure 11b). The probability in each rank came from the verification rank histogram constructed when investigating systematic error (Figure 10). For example, the probability of the verification occurring in rank #1 is the observed number of rank 1 occurrences divided by the total number of verifications, $5735 / 42065 \approx 0.14$. In assigning probability to the four PQPF categories, each category sums probabilities from the same verification rank numbers as with the uniform ranks method. Notice that in Figure 11a & b that the four dashed lines within the verification ranks are identically placed. Compared to the uniform ranks method, the end result of the weighted ranks method is that the total probability in each category is quite different because, while the same ranks get summed for each category, the probability in each rank is far from uniform.

The ranks' individual probabilities in the weighted ranks method are based on how the ensemble typically performs. The verification rank histogram in this case (Figure 10) showed that in the past the ensemble often under or overforecast *pcp24*. If the process which generated these results is stationary, this information can be used to adjust the probability which gets assigned in each category for future PQPF. By using rank probability based on past performance of the ensemble, systematic errors in the ensemble are compensated for.

Figure 13 summarizes the contrasting PQPF for the sample ensemble forecast, *ENS*, determined by the three different methods. Although the uniform ranks method is a slightly better approximation than the democratic voting method, it can be seen that the difference is not significant. Noticeably different PQPF is produced by the weighted

ranks methods. It will be shown in the next chapter that the calibrated PQPF produced with the weighted ranks method does indeed produce the highest quality forecasts.

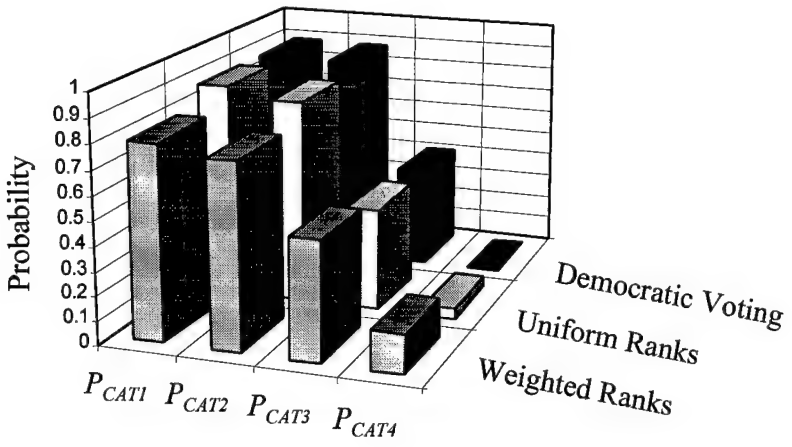


Figure 13. PQPF by category for the sample ensemble forecast, *ENS*, determined by the three different methods.

4. *Analysis and Results*

a. Overview

This chapter presents the findings of this research. Since this research is so novel, it was necessary to make many choices concerning how best to construct the calibration for PQPF. Section b gives details of how the calibration was designed and justifies the major choices that were made. Next, section c presents and discusses the results of the application of the calibration. The accuracy and reliability of the calibrated PQPF produced from the weighted ranks method is compared against PQPF from the democratic voting method, uniform ranks method, and persistence forecasts. Section d then presents an analysis of the limits of predictability of cumulative precipitation. Lastly, section e is an aside on the subject of the difficulty of probabilistic forecasting.

b. Construction of the Calibration

1) Use of Correlated Data

As described in chapter 3, the MRF *pcp24* ensemble forecasts had to be examined for systematic errors in order to build a calibration following the weighted ranks method. This was done with the use of verification rank histograms. A histogram testing as nonuniform indicates the presence of systematic error in the ensemble forecast. For the construction of a histogram, it was briefly noted that samples should be independent (i.e., no correlation between verification points). If correlated data were used, population in the ranks could become unrepresentative. However, it was hypothesized that use of

correlated data presented no such problem. If sampling were done over a very disperse geographical region, the unrepresentativeness of the ranks might be balanced out.

Precipitation data is highly correlated over small spatial scales and short time scales. Since data points in the sample space of this research were on average 200 km and 1 day apart, a high degree of correlation was considered likely. The initial question was: How far apart in space and time do samples from the full sample space need to be taken to reach an acceptably low level of correlation? To answer that question, a detailed correlation study of the precipitation data would then have been necessary.

An alternative approach was to simply run the verification ranking process twice, sampling once from the highly correlated data and again from data with reduced correlation (Toth, 1997). If the resulting histograms were to show no statistical difference, it could be confidently concluded that use of correlated data is legitimate in making verification rank histograms.

This hypothesis was tested using ensemble forecasts with 5.5-day lead times. One verification rank histogram was created using the complete sample space (all case days, all grid points), a total of 41704 verifications (Figure 14a). The sample space was then divided up to produce two more histograms from data with lower correlation. Forecast case days were divided by every other day giving two distinct sets of forecast cases (called *A* and *B*) with at least two days (more for missing days) between forecast initial times. The grid (Figure 5a) was divided up in checkerboard fashion giving two distinct sets of grid points (called *a* and *b*) with increased distance between points. One verification rank histogram used data from *A* with *a* (Figure 14b) while the other used *B*

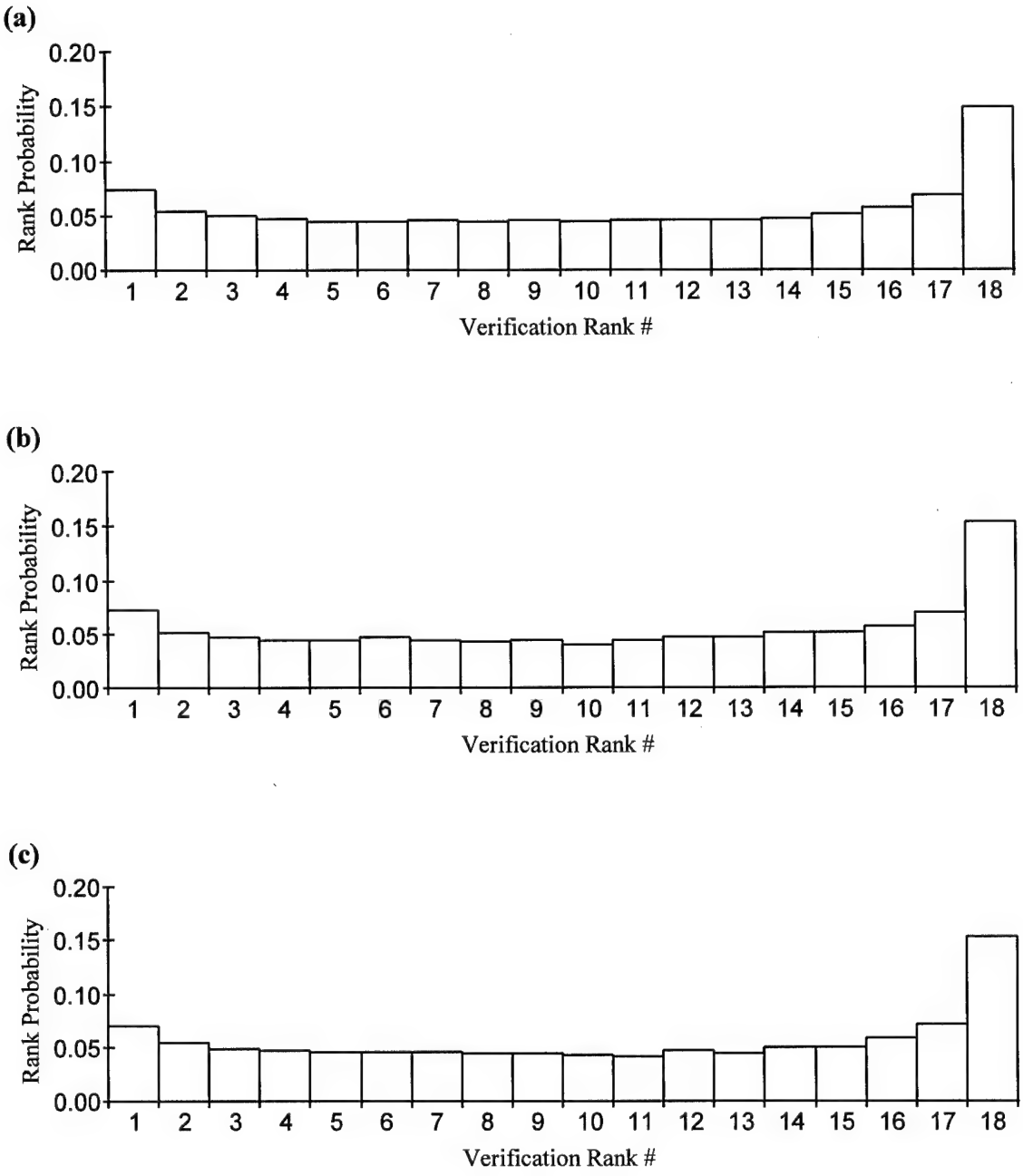


Figure 14. Verification rank histograms of 5.5-day forecasts of the MRF ensemble from (a) highly correlated data using complete sample space, (b) less correlated data using forecasts *A* and grid points *a*, and (c) less correlated data using forecasts *B* and grid points *b*. Minor differences are noticeable, but the histograms are effectively the same.

with *b* (Figure 14c). The size of the sample space for these histograms was 10634 verifications, ¼ of the complete sample space.

From a visual examination of Figure 14, it is difficult to see any differences at all between the three histograms. To be thorough, χ^2 tests were done to test the statistical significance of the differences in the histograms. All comparisons strongly passed with *p* values of 1.0. Tests performed at other forecast lead times gave similar results.

Since verification rank histograms produced at different levels of data correlation yielded the same results, it was concluded that use of correlated data is legitimate. It will become more evident in the following sections of this chapter that this was a very important finding for this research. To produce a robust calibration, the samples had to be finely divided up into groups of similar behavior. If, to achieve low correlation, these divisions had to be made from an already reduced sample space, there would not have been enough data to accomplish the desired calibration. A more generic calibration would still have been possible however.

2) Stationarity of Systematic Errors.

As previously discussed in chapter 3, the most difficult aspect of designing a calibration was in identifying conditions that repeatedly led to similar errors. If the ensemble's systematic errors lacked this stationarity, any attempt at constructing a calibration would be futile. It seems reasonable though to expect the differences between the dynamical system of the model and the atmosphere to be fairly constant, resulting in reoccurring errors in the ensemble. Evidence of this is that numerous technical reports are written on model performance to aid forecasters in interpreting future model output.

Given only the ensemble *pcp24* forecasts, the question was: What conditions can be identified that will result in similar errors? It would be possible to construct one verification rank histogram from the full sample space of all forecasts, at all grid points, and for all 15 valid times. This would make a generic calibration which would improve PQPF for some cases and worsen it for others. What was needed was a flexible calibration which would change according to the conditions of the particular forecast.

The first logical step was to construct a separate histogram for each forecast lead time of the ensemble, fifteen in all. Amount of time into the future was therefore considered a condition upon which the systematic error has some dependence. The reasoning behind this choice is that the character of the ensemble changes over the forecast period. Recall that the IC of the seventeen members are very similar, but since they exist near the maximum growing mode, the members diverge exponentially as the forecasts evolve. This results in an increase in ensemble variability further into the forecast period.

This phenomenon can be seen in the *spaghetti diagrams* of Figure 15. Spaghetti diagrams are one of the tools used in displaying the overwhelming information that comes out of ensemble forecasting. A single isopleth of the parameter is chosen and then plotted on the same chart for each ensemble member. The result looks similar to a dish of pasta, albeit without the sauce or meatballs. Such a diagram is useful in determining confidence in a forecast. When the spaghetti is stuck together, the ensemble members are in strong agreement and a forecast can be made with high confidence. As the spaghetti gets more dispersed and tangled, forecasts are made with decreasing confidence.

(a) Observed pcp_{24} for 12Z, 27 SEP 96

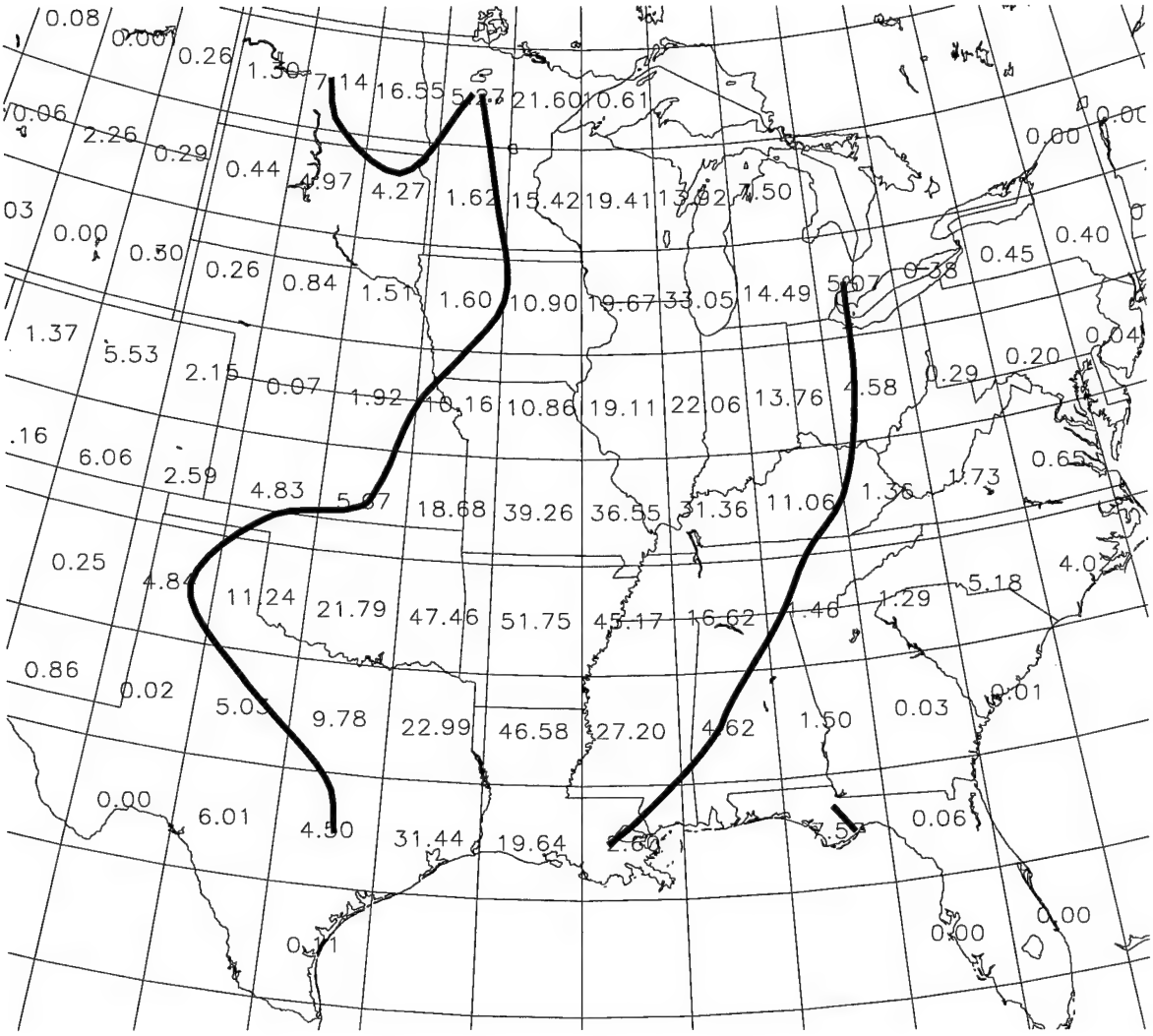


Figure 15. Spaghetti diagrams showing the increase in ensemble variability with increase of forecast lead time. Five different ensemble forecasts of the same event are displayed. Panel (a) shows observed $pcp24$ for 12Z, 27 SEP 96 with a computer analysis of the 6.35-mm isohyet. Panels (b) through (f) are 1.5-day through 5.5-day forecasts, each showing seventeen possible locations for the 6.35-mm isohyet. Initial times of these forecasts were 00Z on (b) 26 SEP 96, (c) 25 SEP 96, (d) 24 SEP 96, (e) 23 SEP 96, and (f) 22 SEP 96.

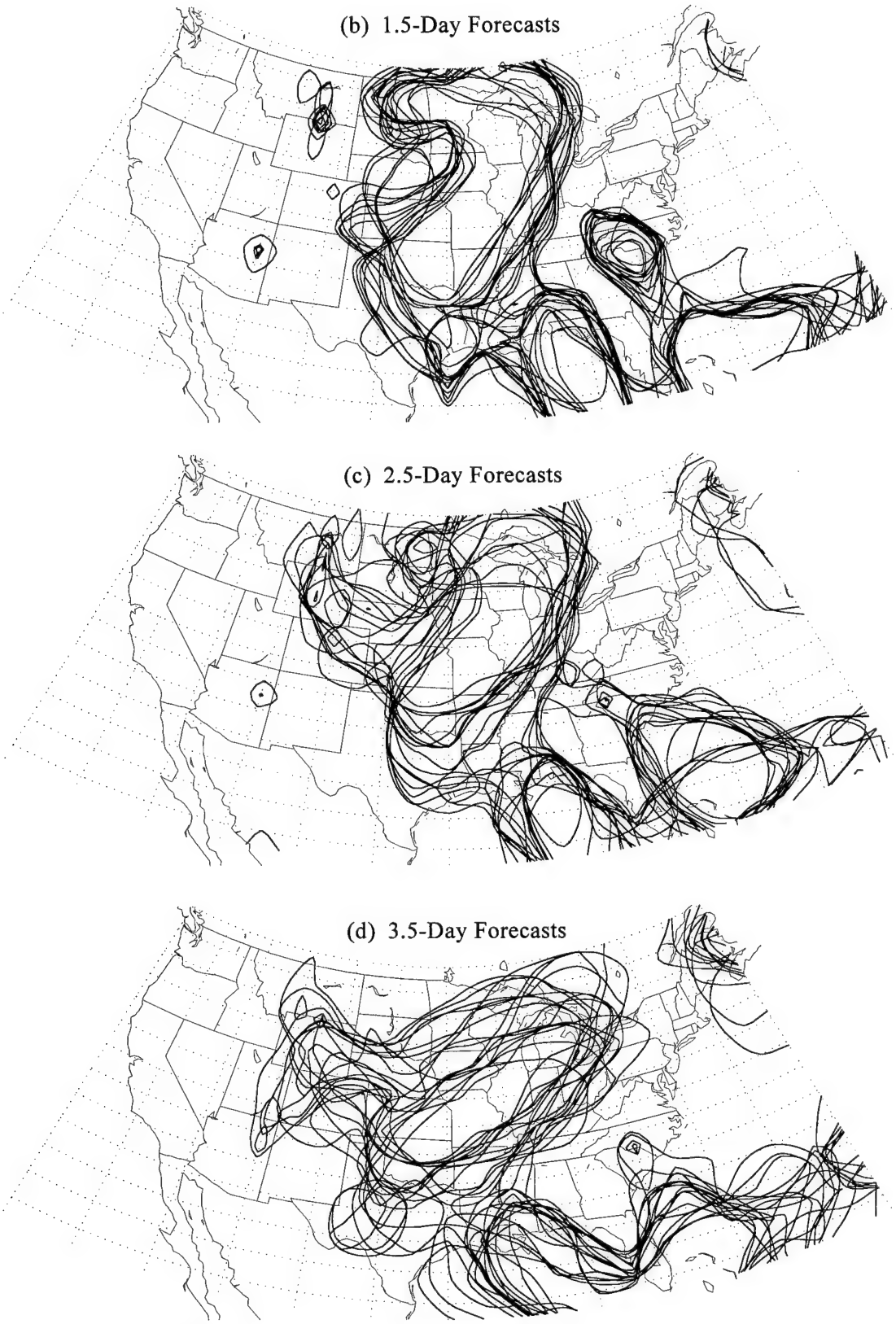


Figure 15. (continued)

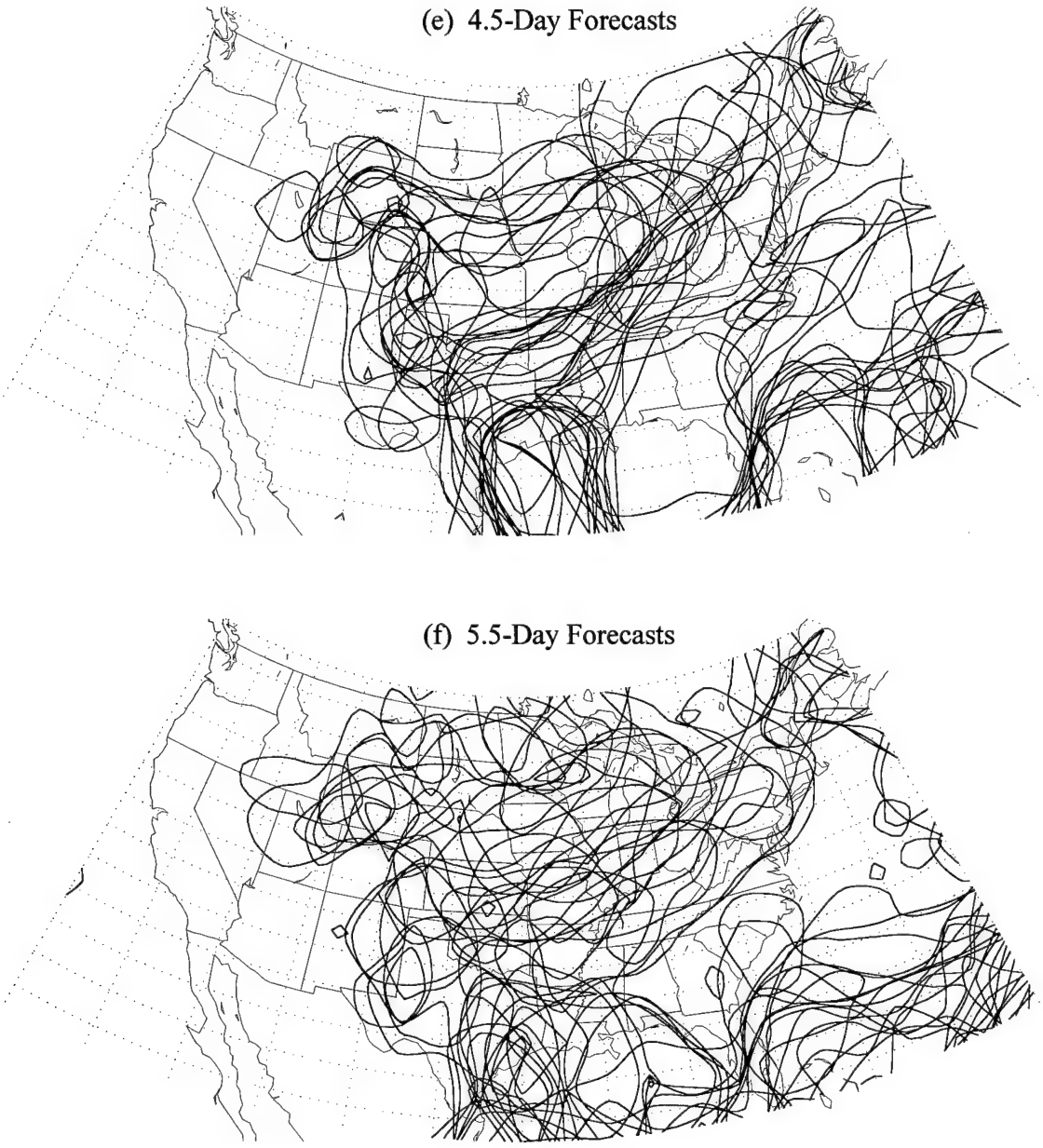


Figure 15. (continued)

In Figure 15, the chosen isohyet (line of constant precipitation amount) was *pcp24* of 6.35 mm, the category 2 threshold. Five separate ensemble forecasts made in the five days leading up to 27 SEP 96 are shown in (b) through (f). For comparison to the verification, the grid of observed values of *pcp24* for 12Z on 27 SEP 96 is displayed in (a) along with the computer analysis of the 6.35-mm isohyet. The prognosis of all five ensemble forecasts is roughly the same, significant rainfall from the Midwest to the Gulf of Mexico. What is important to notice is that the spread of the members increases with forecast lead time. This is the direct result of the divergence of members' trajectories in phase space and has an impact on systematic error.

To explore how the systematic errors change over the forecast period, verification rank histograms were constructed for each forecast lead time. The resulting histograms, converted to probability for better comparison, are shown for a few of the forecast lead times (Figure 16). Comparatively low ensemble variability in the early part of the forecast period appears to increase the chance for the verification to occur in the outer ranks. As ensemble variability increases, the verification seems to occur more often in the inner ranks. This result follows intuition but is misleading because these histograms are still much too general. A more in-depth analysis was required.

Since increased spread of the ensemble over time affected the relationship between the ensemble members and the verification, the next step was to investigate the impact of increased spread over space. This was the approach used by Hamill and Colucci (1997) in designing a calibration for short range ensemble forecasts. Ensemble variability is quantified by the sample standard deviation (s) of the seventeen ensemble

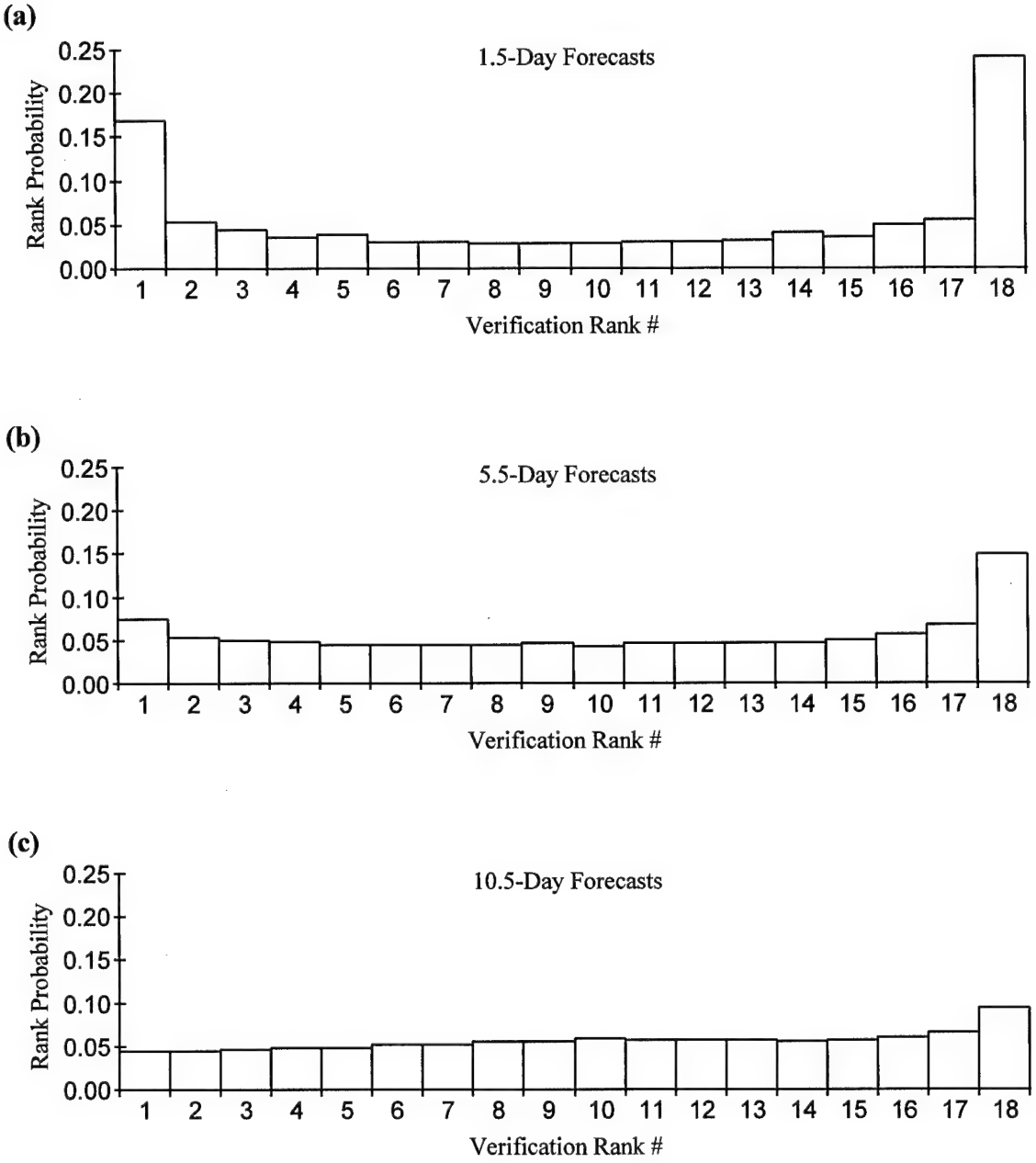


Figure 16. Verification rank histograms at forecast lead times (a) 1.5-day, (b) 5.5-day, and (c) 10.5-day using the full sample space at each time. All histograms test as nonuniform, but there is a progression toward uniform probability as forecast lead time increases. This is due to the increase in ensemble variability further into the forecast period.

members at a grid point. In an ensemble forecast at one valid time, s varies significantly over the grid. Generally speaking, the wetter the forecast at a point, the higher the value of s.

The next question was how to divide up the range of s in constructing multiple verification rank histograms at one forecast lead time. Hamill and Colucci (1997) constructed only three verification rank histograms based on high, medium, and low s, since they were limited by a small number of samples. Since this research had a plethora of data, a more detailed analysis was possible.

The distribution of s for wet forecasts (at least one member forecasts some precipitation) was found to be skewed to the right with a minimum of 0.024 mm. Totally dry forecasts of course have 0.0 for a s value. With increased forecast lead time, the s distribution steadily shifts away from the origin becoming more normal (Figure 17).

With roughly 20000 sample verifications available in the training data set at each lead time, it was possible to divide s into many class intervals while at the same time maintaining large enough subsample space sizes. A unique verification rank histogram could then be constructed for each interval. Sixteen class intervals were chosen based on the distribution of s to give around 1000 samples for each histogram. The range of the class intervals therefore increases with increasing s. Table 3 gives the class intervals used in constructing the 16 histograms for 2.5-day forecasts. Note that dry forecasts ($s = 0.0$) are a unique class and were handled quite differently, since they represent a special case.

The question for dry forecasts was: When all members forecast 0.0 mm, what is the chance of getting some precipitation and how much? In other words, should dry

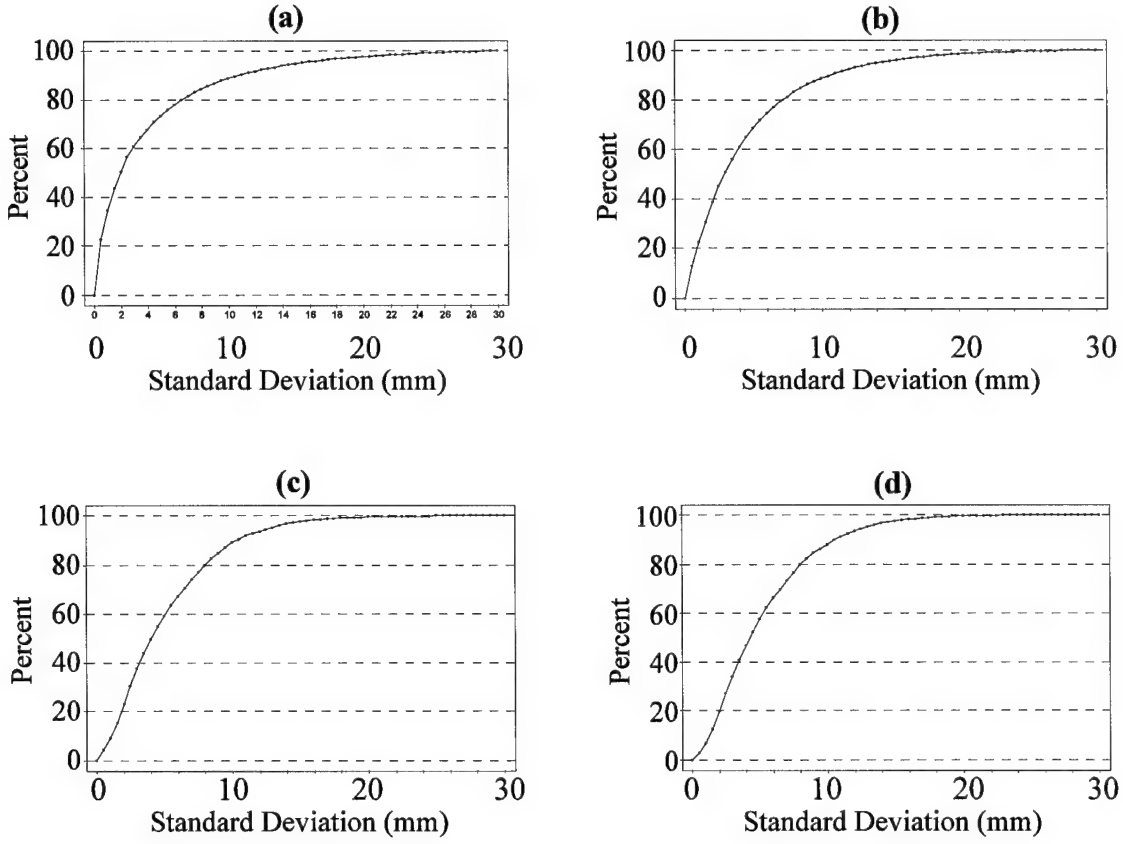


Figure 17. Empirical cumulative distribution of sample ensemble standard deviation from 21324 wet cases of (a) 1.5-day forecasts, (b) 5.5-day forecasts, (c) 10.5-day forecasts, and (d) 15.5-day forecasts. Note the steady progression toward a normal CDF.

forecasts be figured in to the calibration? Figure 18 gives the answer. In nearly half of the 2.5-day dry forecast cases some precipitation was observed, but the amount of precipitation was very low. This is an indication that the observation data may indeed be an overestimate of precipitation for dry events. The majority of the cases do not reach the category 1 threshold of 2.54 mm, making them insignificant for PQPF. Therefore, for the calibrated PQPF, each category is assigned 0% probability in the case of a dry forecast.

Table 3. Class intervals of ensemble standard deviation for 2.5-day forecasts.

Interval Number	Sample Size	Range of s	Median of Interval	ln of Median
N/A	4089	= 0.0	N/A	N/A
1	1504	$0.0 \leq \ldots < 0.1$	0.05	-3.00
2	1456	$0.1 \leq \ldots < 0.3$	0.20	-1.61
3	1043	$0.3 \leq \ldots < 0.5$	0.40	-0.92
4	915	$0.5 \leq \ldots < 0.7$	0.60	-0.51
5	1122	$0.7 \leq \ldots < 1.0$	0.85	-0.16
6	899	$1.0 \leq \ldots < 1.3$	1.15	0.14
7	1084	$1.3 \leq \ldots < 1.7$	1.50	0.41
8	941	$1.7 \leq \ldots < 2.1$	1.90	0.64
9	1010	$2.1 \leq \ldots < 2.6$	2.35	0.85
10	1001	$2.6 \leq \ldots < 3.2$	2.90	1.06
11	1214	$3.2 \leq \ldots < 4.0$	3.60	1.28
12	1194	$4.0 \leq \ldots < 5.0$	4.50	1.50
13	1243	$5.0 \leq \ldots < 6.3$	5.65	1.73
14	1287	$6.3 \leq \ldots < 8.2$	7.25	1.98
15	1205	$8.2 \leq \ldots < 11.5$	9.85	2.29
16	1120	$11.5 \leq \ldots < 25.0$	18.25	2.90

Using the class intervals in Table 3, the resulting 16 verification rank histograms gave a very interesting picture when viewed together (Figure 19a). Discounting the noise due to the reduced sample sizes, the rank probability appeared to vary smoothly along the class intervals as well as across the ranks. This result is actually what should be expected if the systematic error does have some dependence on ensemble variability. The consequence of this discovery is that it was possible to fit functions to the behavior at each rank. Instead of a discrete number of verification rank histograms at each lead time, there could now be an infinite number.

Before attempting to find these functions, the histograms at each class interval had to be smoothed to remove the noise between the ranks. The first step of the smoothing

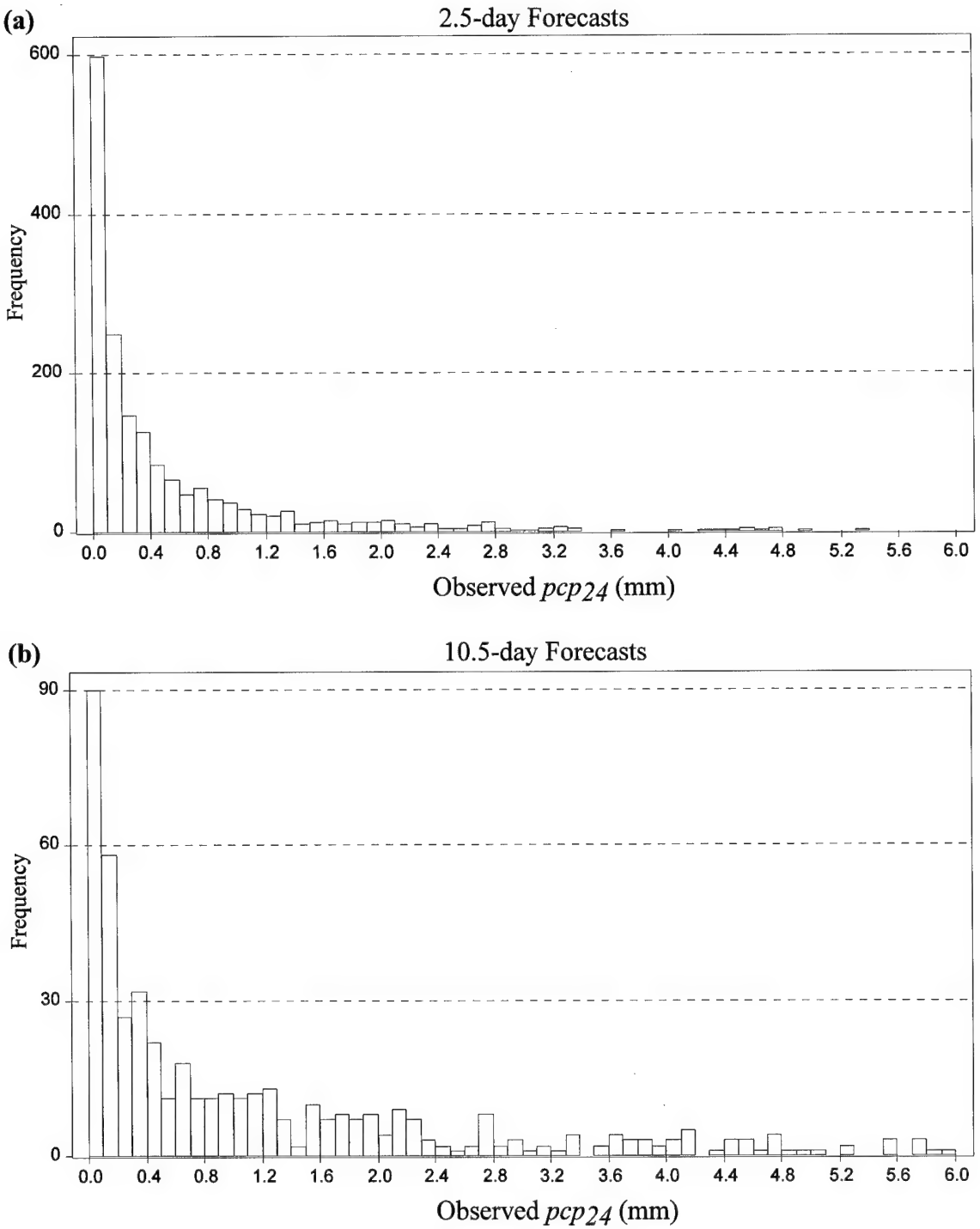


Figure 18. Histograms of observed *pcp24* for dry ensemble forecasts (all 17 members forecast 0.0 mm) with class interval of 0.1 mm for (a) 1756 cases of 2.5-day forecasts, and (b) 473 cases of 10.5-day forecasts. Most of the observations are very close to zero, and only rarely does an observation occur above the category 1 threshold (2.54 mm).

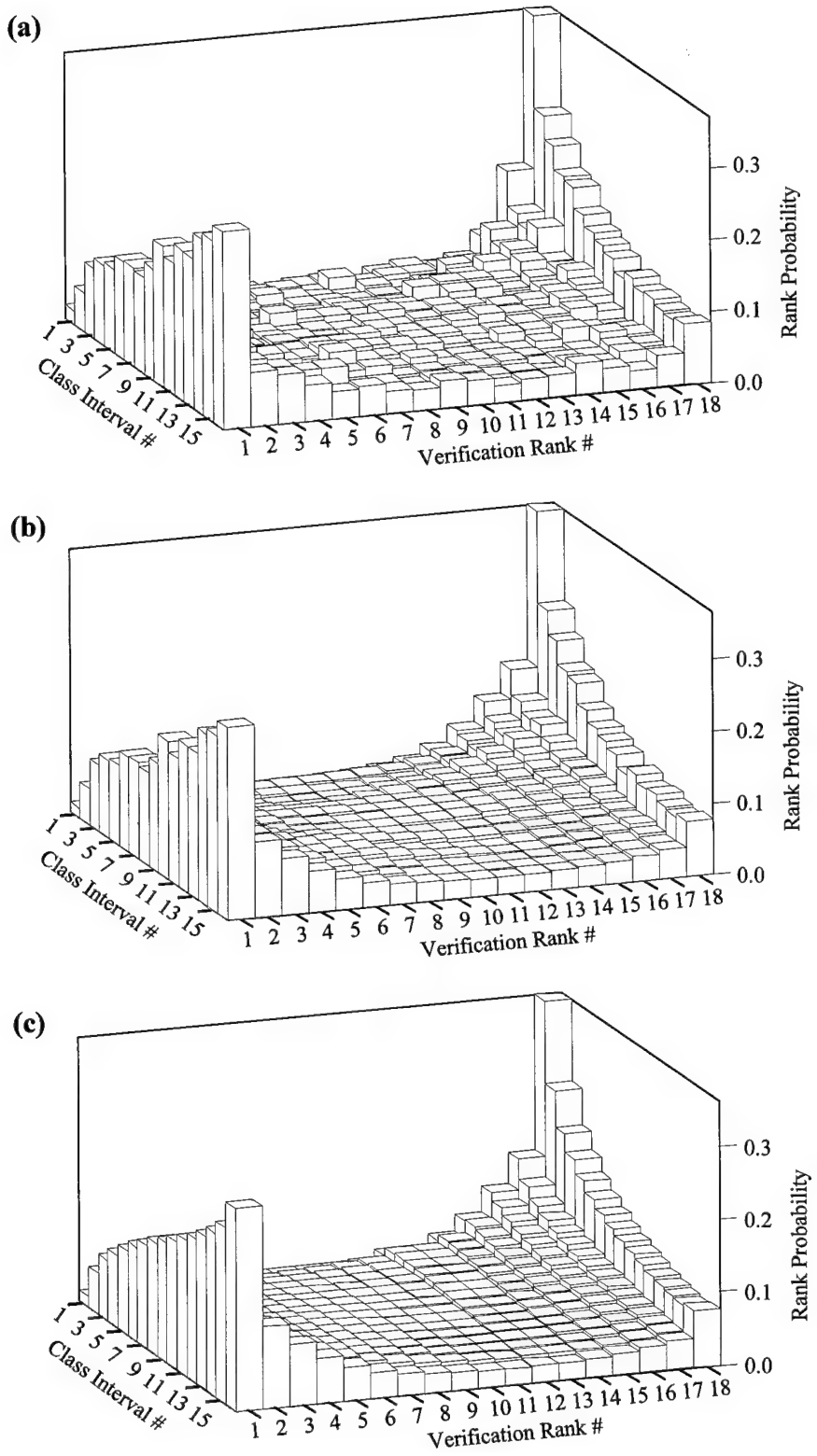


Figure 19. Processing of verification rank histograms for 2.5-day forecasts. (a) Raw data with noise in both directions. (b) After smoothing between the ranks at each class interval of standard deviation. (c) After fitting third-order polynomials at each rank.

process was accomplished with the Mathcad software function called *ksmooth* using a bandwidth of 5. The smoothed value of rank probability (*RP*) for rank *i* was found with Equation 7 (adapted from Equation on p 312, Mathsoft, 1995).

$$RP_i = \frac{\sum_{j=1}^{18} K\left(\frac{r_i - r_j}{5}\right) RP_j}{\sum_{j=1}^{18} K\left(\frac{r_i - r_j}{5}\right)} \tag{7}$$

where *j* is the index for the rank number, *K* is a Gaussian kernel determined by the Mathcad software, and *r* is the rank number. Next the probability of the outside ranks (#1 and #18) were reset to their pre-smoothed value. This was done because the smoothing function reduced the probability in these ranks too much. Lastly, the histograms were normalized to one, since the probability over all the ranks must sum to one. Figure 20 shows the raw and processed data for class interval #3 of 2.5-day forecasts. Figure 19b shows all 16 smoothed verification rank histograms.

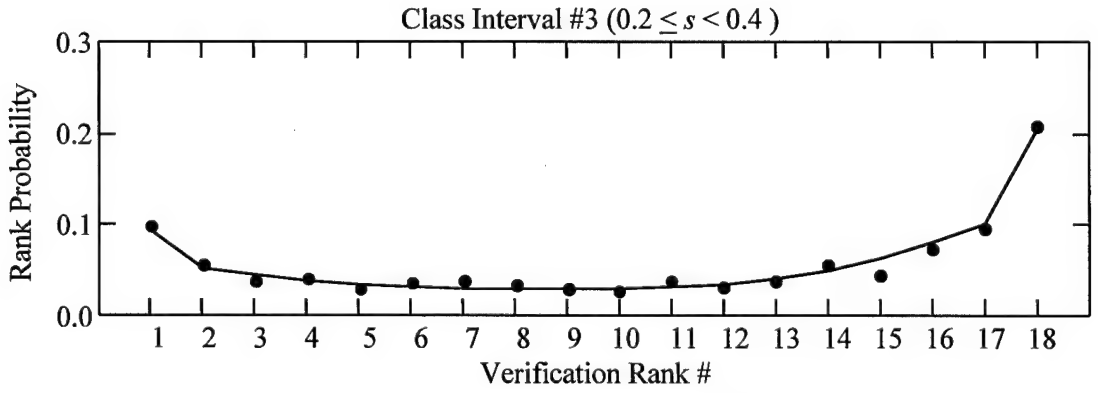


Figure 20. Example of data smoothing of verification rank histogram for class interval #3 of 2.5-day forecasts. Raw data is plotted as dots (•). Solid line connects the values of smoothed, normalized data points.

The next process, also performed in Mathcad, was to fit a function at each rank. First, the median of each class interval was chosen to represent the value of the independent variable (standard deviation) for the corresponding 18 values of the dependent variable (rank probability). Table 3 back on page 58 gives the median values for the 2.5-day forecasts. Plotted on linear scale, the data showed a logarithmic quality so the independent variable was transformed to a log scale. It was then found that third-order polynomials made an excellent fit to all the ranks. Figure 21 is an example of this process for the function fit to rank #17 for 2.5-day forecasts.

For the sake of comparison, Figure 19c displays the resulting histograms for all the original class intervals of the 2.5-day forecasts. The information now available should no longer be displayed in this fashion since any one of the multitude of possible histograms does not represent a discrete range of standard deviation anymore. There is now a continuum where any particular value of standard deviation determines a histogram with a unique set of 18 rank probabilities.

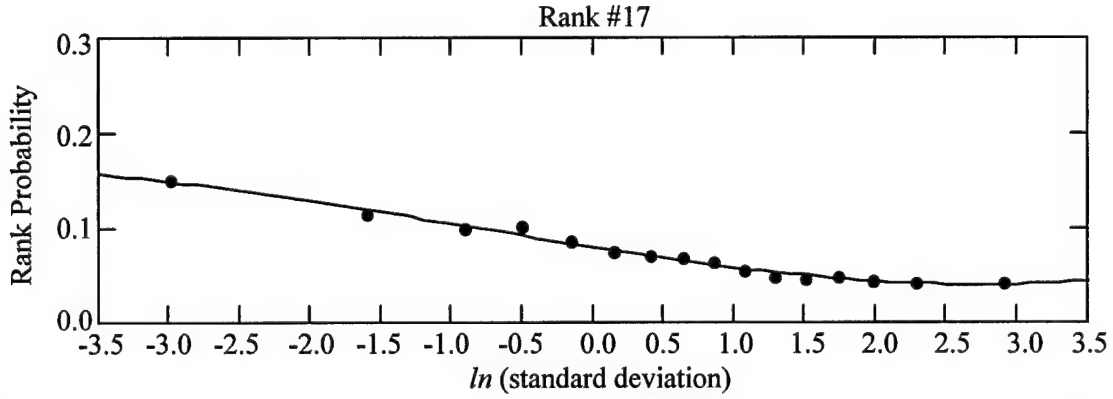


Figure 21. The rank #17 third-order polynomial fit to the natural log transformation of the median values of the 16 class intervals. Raw data is plotted as dots (•).

The name given to this continuum of verification rank histograms is a *probability surface*. An example for 2.5-day forecasts is shown in Figure 22 along with the coefficients of the third-order polynomials at each rank in Table 4. Technically, it is not a surface because while it is continuous along the ln(s) axis, the ranks are still discrete integer quantities. The surface is really just a series of 18 separate functions. Viewing the functions as a surface is useful tool for interpreting and understanding how they work in the calibration scheme.

Notice that the ln(s) scale extends beyond the range of the transformed median values in both directions. For these values, the functions are extrapolating to ln(s) values outside of the original data values so the rank distribution becomes less realistic. The lowest possible value of s is 0.024 mm, which transforms to –3.73 on the log scale. The high value plotted of 4.0 is for $s \approx 55$ mm, an extremely high s value which rarely occurs. While the extrapolated regions of the probability surface may be less realistic, this does not have a serious impact on the calibration since the majority of s values will not occur there. On some surfaces, the extrapolated curves fall below 0.0 probability. For these cases, the rank probabilities are frozen at the last s value that had no probability values less than 0.0. All s values beyond this point then repeat the same histogram.

As previously discussed, verification rank histograms constructed at various lead times showed noticeable differences. It was theorized that this difference was due mainly to increase in ensemble spread (standard deviation) over the forecast period. Since the independent variable of a probability surface is also standard deviation, it is logical to believe that one probability surface could be constructed to include all forecast lead

Table 4. Coefficients of the third-order polynomials for each rank of 2.5-day forecasts.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.1224011069	0.0462129592	0.0030028403	-0.0001408405
2	0.0629425734	0.0191124113	0.0003974715	-0.0003835922
3	0.0523940448	0.0148657563	0.0000405033	-0.0004423607
4	0.0438376060	0.0114929703	-0.0002944019	-0.0004983181
5	0.0382419099	0.0091084098	-0.0006195941	-0.0005342973
6	0.0353169303	0.0073305730	-0.0009218651	-0.0005249800
7	0.0339843883	0.0056719217	-0.0011547687	-0.0004508821
8	0.0333723629	0.0039249373	-0.0012895197	-0.0003182181
9	0.0332651455	0.0021737302	-0.0013599443	-0.0001610409
10	0.0338671813	0.0005634714	-0.0014328944	-0.0000161406
11	0.0354209811	-0.0009108059	-0.0015364163	0.0001057972
12	0.0380993826	-0.0024602339	-0.0016327575	0.0002199798
13	0.0421687845	-0.0044560885	-0.0016339727	0.0003388527
14	0.0481424036	-0.0073931691	-0.0014015055	0.0004533682
15	0.0565541426	-0.0117156145	-0.0007764539	0.0005454445
16	0.0673212360	-0.0173883491	0.0002938198	0.0006061855
17	0.0794237736	-0.0237213075	0.0016654037	0.0006375191
18	0.1432460468	-0.0524115719	0.0086540553	0.0005635236

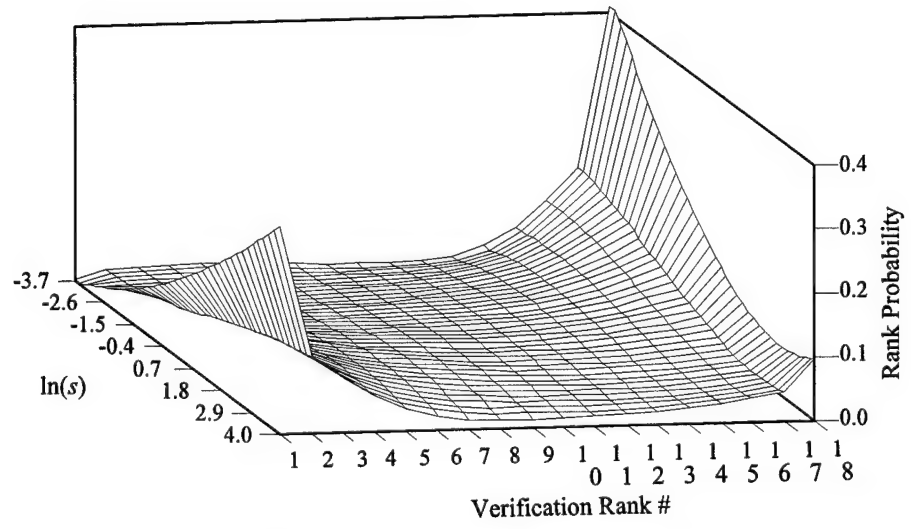


Figure 22. Probability surface for 2.5-day forecasts. The curves at each rank are the third-order polynomials with coefficients given in Table 4.

times. This would eliminate much complication. Unfortunately, this simplification proved too elusive.

Probability surfaces created for different lead times have significant differences. The most noticeable changes occur within the first six days of the forecast period (Figure 23a - c). For forecasts beyond 5.5 days, the probability surfaces still steadily change with increasing lead time but not as drastically. The fact that there are noticeable changes from one lead time to the next means that a particular value of s, say 0.7 on the log scale, does not correspond to the same systematic error at each lead time. For a 1.5-day ensemble forecast with $s = 0.7$ (Figure 23a), the verification is most likely to occur in rank #1 or #18. Alternatively, the same s value for a 5.5-day forecast (Figure 23c) will most likely have a verification occur in one of the lower ranks or rank #18.

Besides being useful for the calibration, these surfaces convey the MRF's precipitation bias. Consider a very wet (high value of s) forecast with 1.5-day lead time (Figure 23a). Since it is early in the valid period, ensemble members are in relatively good agreement compared to previous forecasts for the same events. One might conclude that the MRF has a serious problem with its parameterization of moisture resulting in an overforecasting bias. More plausibly, high precipitation occurs less often and over a more limited area than low or no precipitation. A wet forecast that is in error in space and/or time will then likely overforecast the verification value. Notice that as the valid time increases, this phenomenon becomes less pronounced because of increasing ensemble divergence. The wet forecasts are still overforecasts, but with a larger spread the ensemble manages to encompass the verification more often.

For producing the calibrated PQPF for this research, the weighted ranks method used this set of fifteen probability surfaces. The third-order polynomial coefficients for all fifteen surfaces are given in appendix A. Further analysis of systematic error was explored to examine the possibility of an even more flexible determination of rank probabilities. The results were however not used in the final calibration.

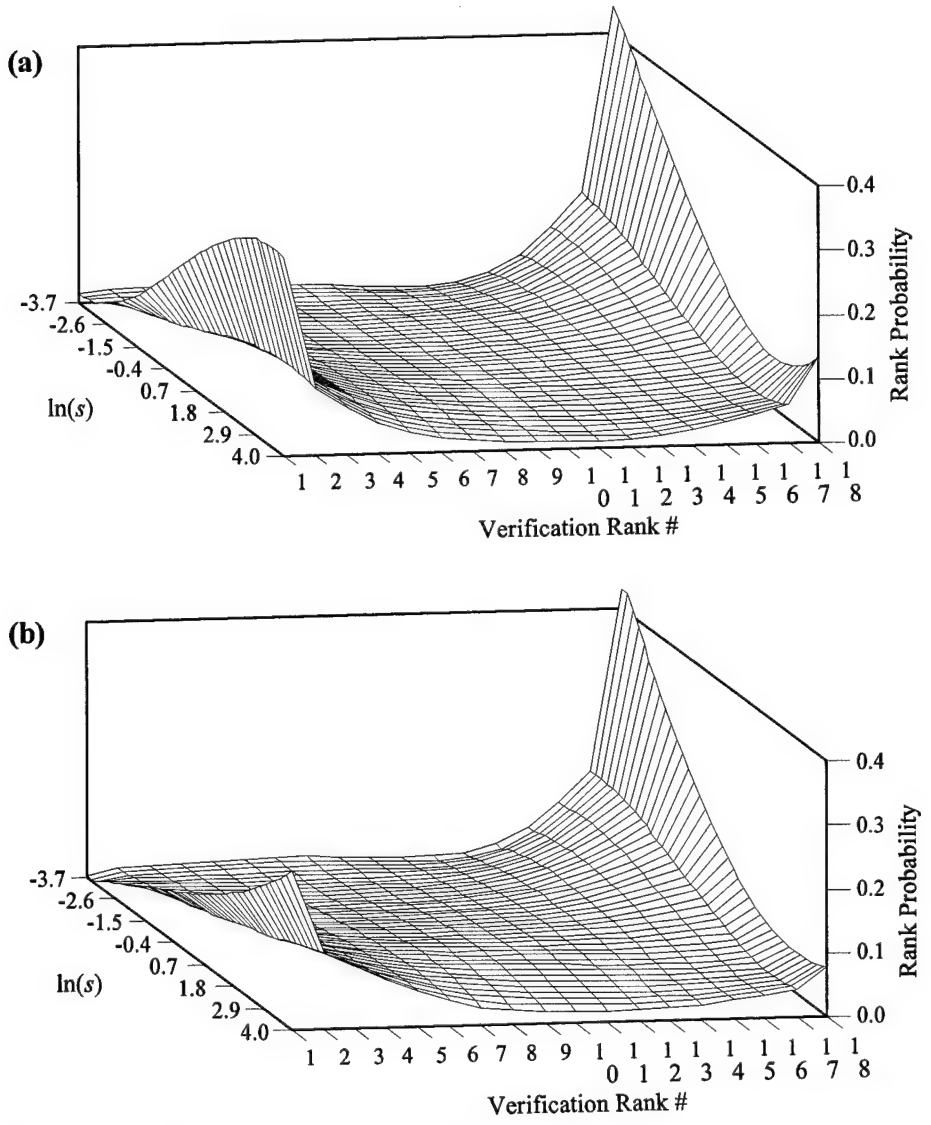


Figure 23. A few of the 15 probability surfaces of the calibration. Surfaces shown are for (a) 1.5-day, (b) 3.5-day, (c) 5.5-day, (d) 10.5-day, and (e) 15.5-day forecasts.

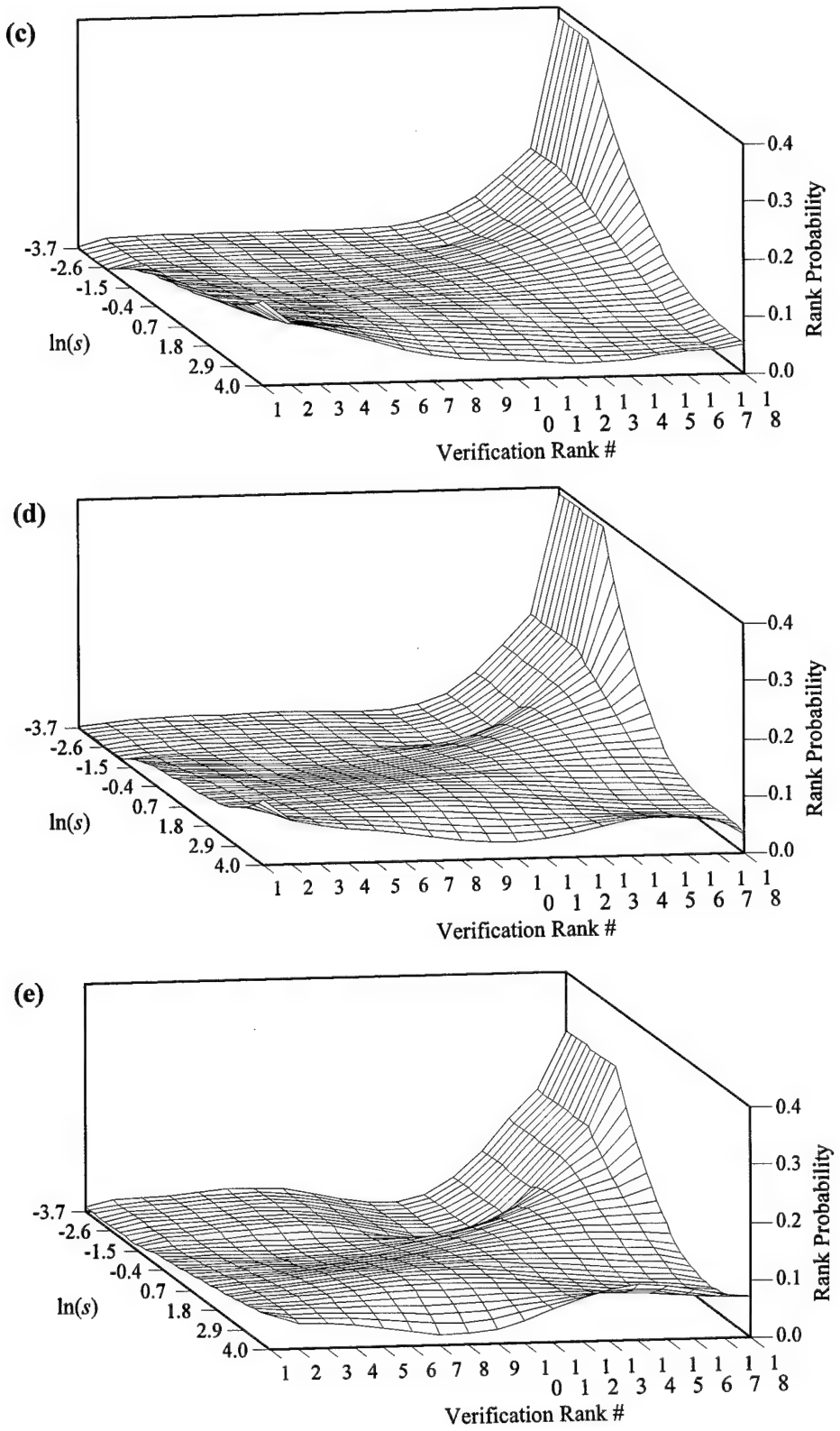


Figure 23. (continued)

3) Regime Dependence

The verification rank distribution was shown to have a strong dependence on both ensemble standard deviation and forecast lead time. This indicates that the strategy of using the probability surfaces should produce a good calibration. Is there a further stratification that may provide a better calibration? To examine this possibility, the *regime* dependence of the surfaces was briefly explored.

A weather regime is described by a defined set of conditions over an area in which some similar relationships and/or influences on atmospheric dynamics are shared (i.e., low atmospheric stability, warm ocean temperatures, continental polar air, etc.). The reasoning behind a regime dependence of systematic error is that within a one weather regime, the ensemble's systematic errors will actually differ from another regime due to differences in dynamical forcing. Each regime would then require a more specific calibration. Since this research had only the ensemble *pcp24* data to work with, regimes defined only by geography and season were explored here.

The approach was to create two separate probability surfaces for 5.5-day forecasts using data from two completely different regimes. If these surfaces proved to be different, then a regime dependence for the surfaces could be concluded. Two geographical regimes were created by dividing the grid diagonally from WA to FL giving a NE US regime and a SW US regime. Two seasonal regimes were defined as summer (MAY through SEP) and winter (NOV through MAR).

Various combinations of these regimes were used to produce 5.5-day forecast probability surfaces. All the resulting surfaces had similarities but noticeable differences.

Figure 24 shows the probability surfaces for the regimes of NE US in summer and SW US in winter. The same general pattern is evident in both, but they do produce different calibrations. There is a regime dependence.

Thus a more specific calibration than the one used in this research is possible. However, because the differences in the regime probability surfaces are small when

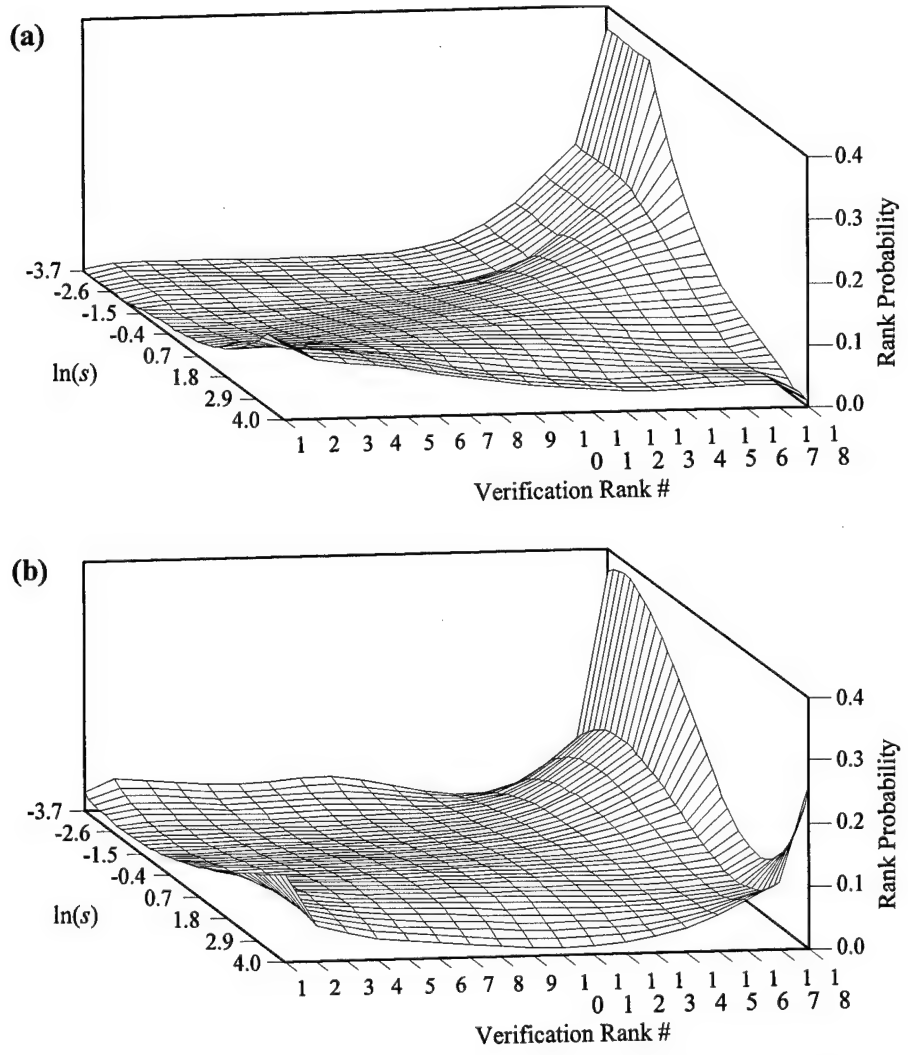


Figure 24. Probability surfaces of 5.5-day forecasts constructed for the different weather regimes of (a) NE US in summer and (b) SW US in winter. The surfaces display the greatest differences in the extreme ranks.

compared with the probability magnitude of the more general 5.5-day surface (Figure 23c), the benefit of this analysis would be small compared to the amount of effort it would take. For this reason, regime dependent probability surfaces were not pursued.

c. PQPF Comparison

1) Value of Improved PQPF

Before examining the quantitative improvements of the calibrated PQPF, it should be shown that the calibration produces a forecast that represents a significant difference to the user (both the forecaster and the customer). Increased quality of a probability forecast is desirable but if the new, improved forecast differs by only a few percent, the improvement might be of no consequence to the user.

As a hypothetical example, say the MRF shows the possibility of a major winter storm five days hence along the East Coast of the US. The wing commander at McGuire AFB, NJ, is concerned about the operational impact of heavy snow. He asks for the probability of getting more than 1 foot of snow over a 24-hour period. Using the general 12 to 1 rule for liquid precipitation to snow amount, the forecaster accesses the probability at 11% by using the uncalibrated CAT4 PQPF ($pcp24 > 25.4$ mm or 1.0") with a 5.5-day lead time. From this forecast, the wing commander's decision is to not prepare to deploy his aircraft to an alternative location.

If the calibrated PQPF produced a slightly higher probability, say, 13%, the wing commander would likely make the same decision, so the improvement would not be useful. Alternatively, suppose the calibrated PQPF gave a CAT4 probability of 24%. Now the commander's decision may be to begin preparation for a deployment. The

difference between uncalibrated and calibrated PQPF may need to be at least 5 – 10% to represent a useful improvement.

Figure 25 gives a typical example of the forecast probability differences between the democratic voting, uniform ranks and weighted ranks methods. Comparing the uniform ranks PQPF (Figure 25d) to democratic voting PQPF (Figure 25c), the differences are quite small. Therefore, even though the uniform ranks PQPF is of higher quality, as will be seen in the next section of this chapter, the improvement is insignificant from an operations point of view.

Conversely, the calibrated PQPF of the weighted ranks method does show significant differences over the democratic voting PQPF (Figure 25a). From west of the Mississippi up through the Ohio Valley, the calibrated PQPF is less than the uncalibrated by up to 20%. According to Figure 25b, the verification failed to reach the category 2 threshold over this region so the calibration corrected PQPF in the right direction. The calibration had the capability to recognize that for the conditions of this forecast, the ensemble is typically an overforecast. Similarly in Michigan and parts of western Virginia and southern West Virginia, the calibrated PQPF predicted a slightly higher chance of occurrence, which was correct.

The calibration does not, however, always make the correct adjustments. Consider eastern Kentucky where probability was decreased but the verification did occur above the threshold. Since the calibration is based on the typical errors of the ensemble, it only makes the right adjustment most of the time. The value of the calibration is that over many samples, its incorrect adjustments are far outweighed by the correct ones.

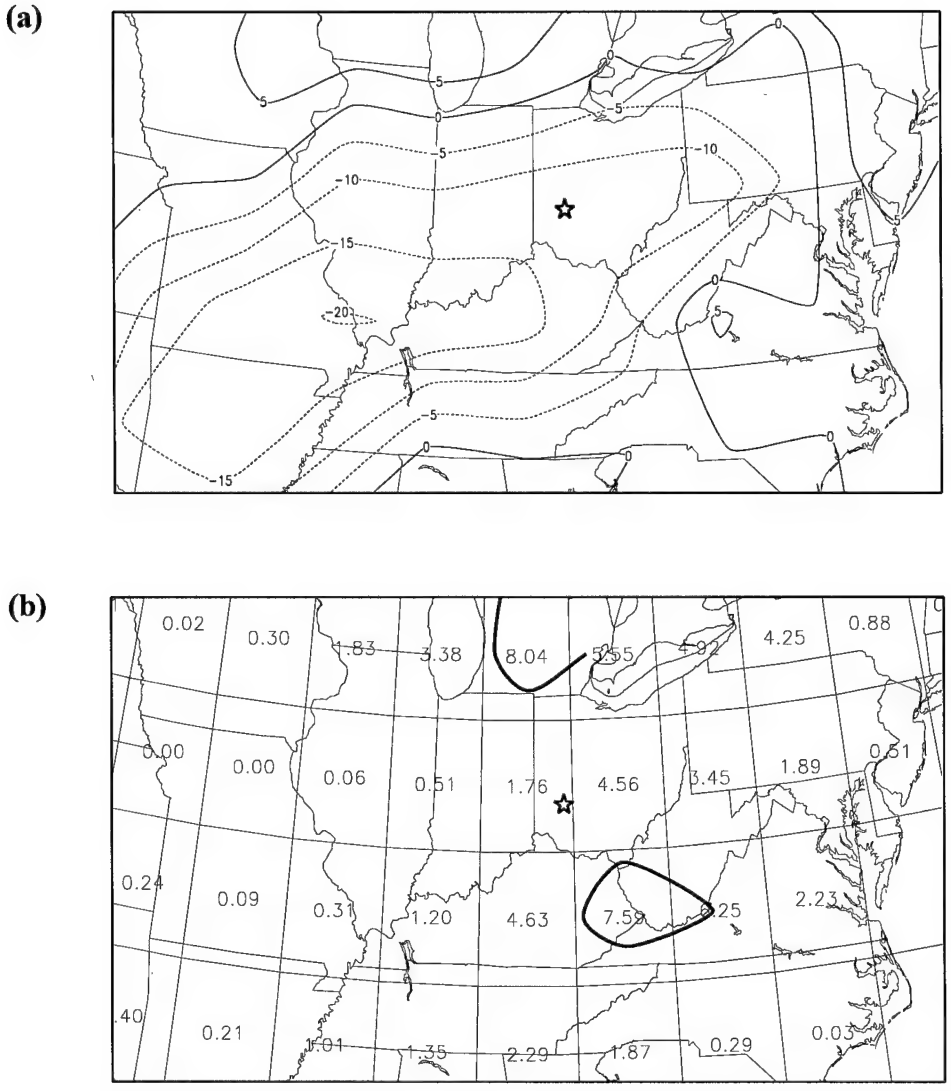
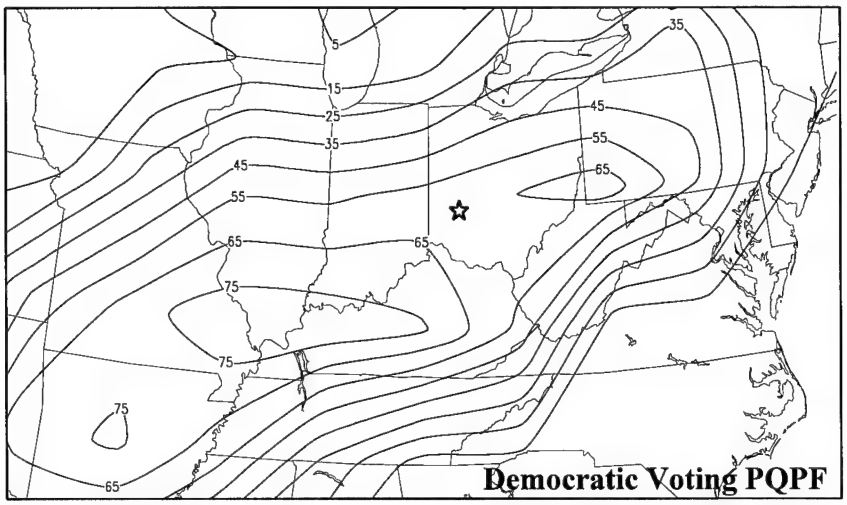
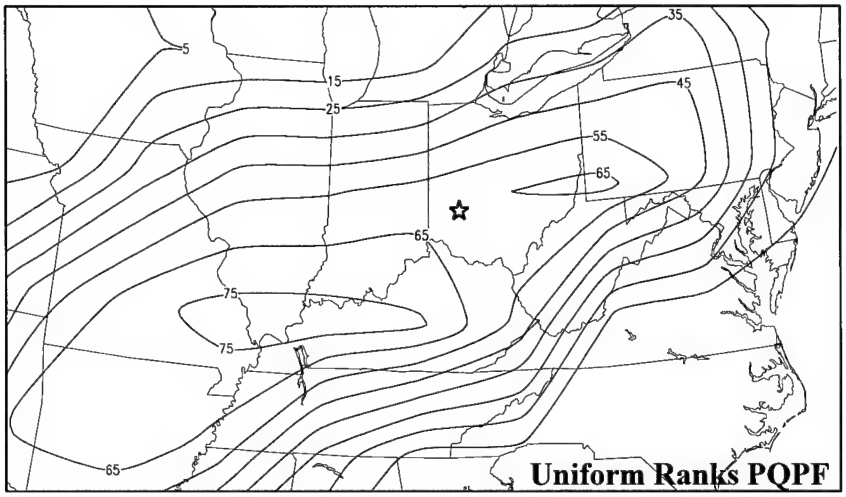


Figure 25. Sample PQPF for CAT2 (pcp_{24} > 6.35 mm) from the MRF ensemble initialized at 00Z on 12 MAY 97. Forecasts with lead time of 3.5 days (i.e., pcp_{24} from 12Z 14 MAY to 12Z 15 MAY) produced by (c) democratic voting method, (d) uniform ranks method, and (e) weighted ranks method. The numerical difference between (c) and (e) is in panel (a). Panel (b) is the observed pcp_{24} with a computer analysis of the 6.35-mm isohyet.

(c)



(d)



(e)

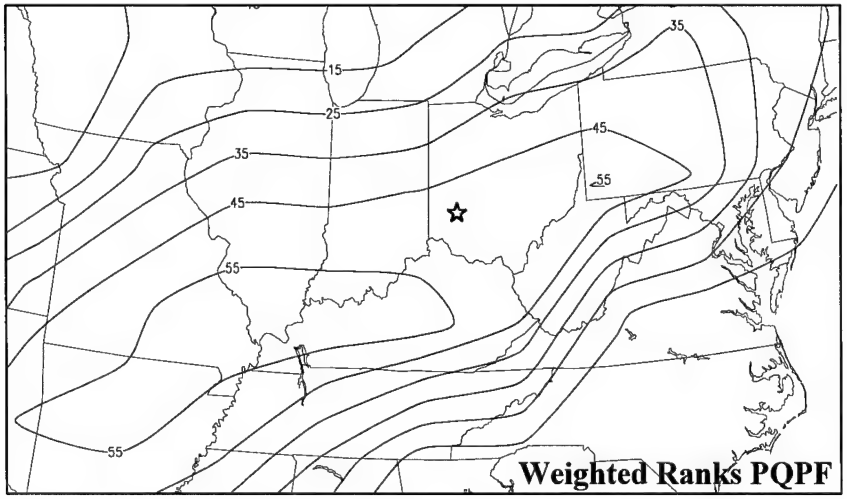


Figure 25. (continued)

There is a cautionary note to be made concerning this analysis of Figure 25. Instances in which the PQPF is nonzero and *pcp24* fails to occur above the threshold do not necessarily denote a bad forecast. Consider the weighted ranks PQPF for Wright Patterson AFB, OH, indicated by the star in Figure 25c. The probability of getting *pcp24* > 6.35 mm was put at 48%, but only 1.76 mm was observed in that grid box. What the probability really means is that given 100 cases with the exact same ensemble IC, the atmosphere will evolve to produce *pcp24* > 6.35 mm at that location in 48 of the cases. Each of these hypothetical 100 cases is an attempt to forecast a very similar but unique trajectory of the atmosphere that yielded the same exact setup in the ensemble. This particular forecast just happened to be for one of the 52 times that *pcp24* did not occur above the threshold.

This example also brings up the point that the calibration can only improve the quality of PQPF to a certain degree (Hamill, 1998). If the model is seriously deficient, post-processing is of little gain. The goal of this research was not to produce more accurate model output but higher quality PQPF based on the typical performance of the MRF ensemble. The calibration improves accuracy and reliability of PQPF by compensating for systematic errors of the MRF ensemble through an interpretation of the raw ensemble forecast data. Improvement in accuracy of the raw ensemble forecast data is obtainable through advancements in the model and/or better ensemble techniques.

2) The Calibration's Improvement of PQPF

This research measured and compared quality of PQPF with four different tools, namely the *Brier score*, *ranked probability score*, *reliability diagram* and the *confidence*

diagram. Explanations of each tool along with their results are presented in that order in this section. The PQPF produced by the weighted ranks method proved to be the best forecasts.

In addition to measuring reliability of the three PQPF methods which were based on the MRF ensemble, a *persistence forecast* was created to compare the reliability of an unskilled forecast. In a persistence weather forecast, whatever weather just happened is forecast to reoccur. (E.g.: If the high temp in Boston was 10°C today, it is forecast to be 10°C tomorrow.)

To create persistence PQPF, the observed pcp_{24} in the 24 hours before the forecast initial time is first converted into categorical probabilities and then used for a set PQPF. For example, say the observed pcp_{24} was 7.2 mm. The event's observed probabilities for the four categories, in vector format, are (1.0, 1.0, 0.0, 0.0) since the verification occurred in CAT1 and CAT2. This is now the PQPF used repeatedly at that grid point for every forecast lead time over the entire 15.5-day valid period. With every grid point handled in this same manner, such a forecast would be quite unskilled.

a) Brier Score

The Brier score (BS), essentially a mean square error measure of a probability forecast, is useful for examining the accuracy of each PQPF category separately (Wilks, 1995). The BS can be calculated for each forecast category using Equation 8 (from Equation 7.22, Wilks, 1995):

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \left(FP_i - OBS_i \right)^2 \qquad (8)$$

where n is the total number of forecasts/observation samples for the category, and FP_i is the forecast probability of the i^{th} sample. OBS_i equals 1 if the value of the predicted variable occurred in the category for the i^{th} sample, and 0 if it did not occur. Therefore, BS varies between 0 (perfectly accurate) and 1 (totally inaccurate).

The tables and figures on the next few pages present the BS for each of the four PQPF categories at each lead time in the forecast period. These results may be considered robust because of the large sample size applied. Sample sizes for each lead time varied within a few hundred of 22000 due to a few missing days of observations. Table 5 through Table 8 give the raw BS for all four forecast types and percent improvement over the democratic voting method for the uniform ranks and weighted ranks methods. Figure 26 through Figure 29 show graphs of the BS.

The most important information here is that PQPF derived using the weighted ranks method is the most accurate since it has the lowest BS. All the BSs generally increase with forecast time and converge indicating two things: (1) PQPF accuracy is lower further into the forecast period which is as anticipated; and (2) the effectiveness of the calibration decreases with forecast time. This point is also evident in the percent improvement over the uncalibrated PQPF shown in Table 5. While quite high early in the forecast period, the improvement steadily falls off. This is due to the decrease in predictability as forecast lead time increases.

Notice that the BS scales on the five graphs are quite different. The range of BS for the lower categories is considerably higher than that of the higher categories. This appears to indicate that PQPF is generally better at higher thresholds – the reason being

Table 5. Brier scores for CAT1 for all lead times.

Lead Time (days)	Democratic Voting BS	Uniform Ranks BS / Improvement		Weighted Ranks BS / Improvement		Persistence BS
1.5	0.1306	0.1268	2.9%	0.1168	10.6%	0.2605
2.5	0.1460	0.1426	2.3%	0.1330	8.9%	0.3239
3.5	0.1599	0.1567	2.0%	0.1489	6.9%	0.3329
4.5	0.1692	0.1665	1.6%	0.1618	4.4%	0.3340
5.5	0.1802	0.1777	1.4%	0.1749	2.9%	0.3316
6.5	0.1906	0.1879	1.4%	0.1854	2.7%	0.3171
7.5	0.1955	0.1930	1.3%	0.1914	2.1%	0.3191
8.5	0.2033	0.2009	1.2%	0.2004	1.4%	0.3394
9.5	0.2033	0.2010	1.1%	0.2018	0.7%	0.3423
10.5	0.2000	0.1981	1.0%	0.1977	1.2%	0.3418
11.5	0.1935	0.1920	0.8%	0.1942	-0.4%	0.3257
12.5	0.1975	0.1960	0.8%	0.1999	-1.2%	0.3358
13.5	0.2008	0.1991	0.8%	0.2034	-1.3%	0.3391
14.5	0.2046	0.2027	0.9%	0.2060	-0.7%	0.3493
15.5	0.2021	0.2003	0.9%	0.2036	-0.7%	0.3426

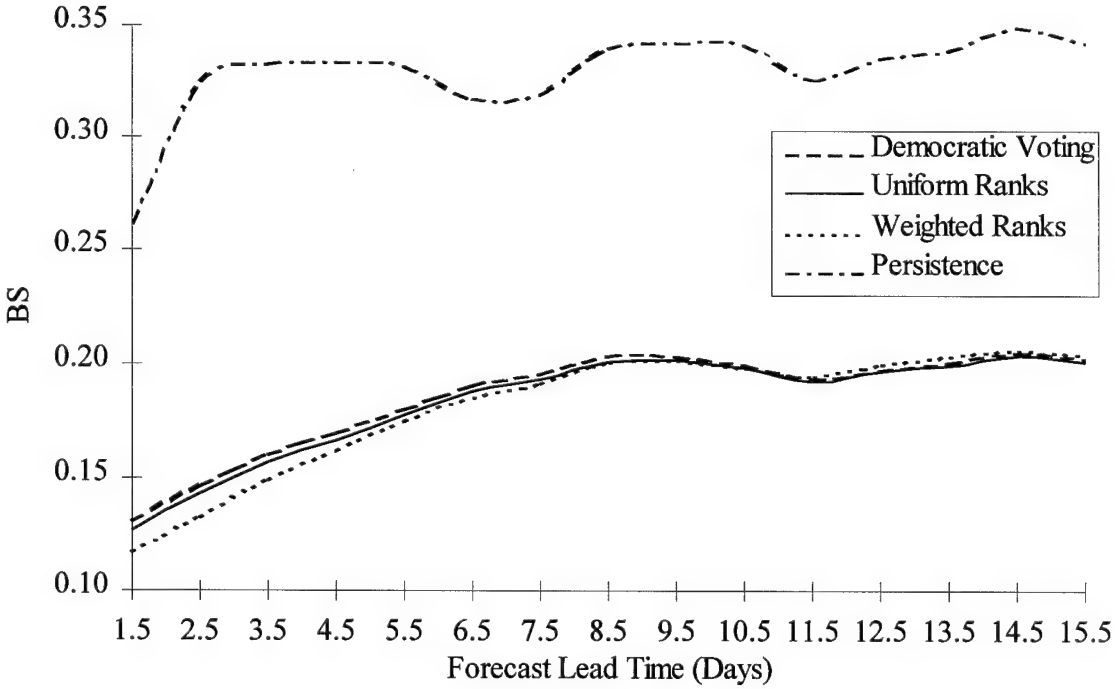


Figure 26. Graph of CAT1 Brier scores over the entire valid period.

Table 6. Brier scores for CAT2 for all lead times.

Lead Time (days)	Democratic Voting BS	Uniform Ranks BS / Improvement		Weighted Ranks BS / Improvement		Persistence BS
1.5	0.0823	0.0796	3.3%	0.0671	18.5%	0.1586
2.5	0.0909	0.0888	2.3%	0.0772	15.1%	0.1866
3.5	0.0970	0.0954	1.6%	0.0866	10.7%	0.1876
4.5	0.1023	0.1010	1.3%	0.0943	7.8%	0.1892
5.5	0.1074	0.1066	0.7%	0.1018	5.2%	0.1889
6.5	0.1094	0.1089	0.5%	0.1044	4.6%	0.1798
7.5	0.1126	0.1122	0.4%	0.1087	3.5%	0.1818
8.5	0.1172	0.1170	0.2%	0.1139	2.8%	0.1918
9.5	0.1142	0.1144	-0.2%	0.1119	2.0%	0.1932
10.5	0.1131	0.1133	-0.2%	0.1095	3.2%	0.1941
11.5	0.1104	0.1108	-0.4%	0.1079	2.3%	0.1843
12.5	0.1121	0.1125	-0.4%	0.1101	1.8%	0.1903
13.5	0.1164	0.1167	-0.3%	0.1150	1.2%	0.1945
14.5	0.1198	0.1200	-0.2%	0.1183	1.3%	0.2022
15.5	0.1183	0.1184	-0.1%	0.1160	1.9%	0.1986

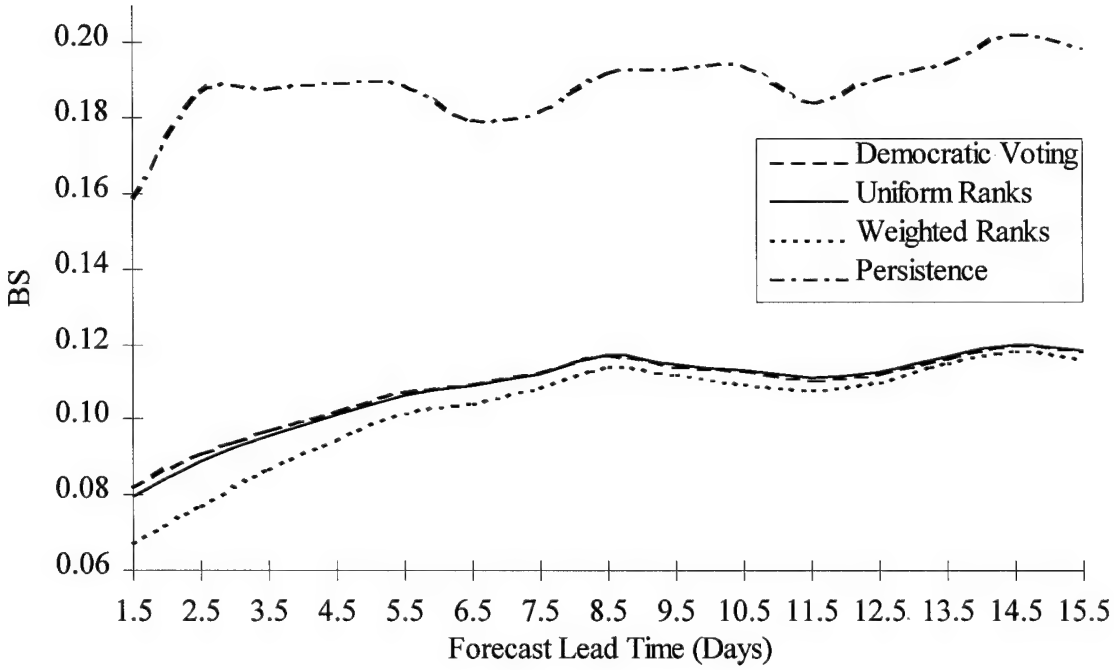


Figure 27. Graph of CAT2 Brier scores over the entire valid period.

Table 7. Brier scores for CAT3 for all lead times.

Lead Time (days)	Democratic Voting BS	Uniform Ranks BS / Improvement		Weighted Ranks BS / Improvement		Persistence BS
1.5	0.0409	0.0399	2.4%	0.0327	20.0%	0.0750
2.5	0.0444	0.0439	1.1%	0.0370	16.7%	0.0847
3.5	0.0449	0.0447	0.4%	0.0397	11.6%	0.0835
4.5	0.0474	0.0473	0.2%	0.0430	9.3%	0.0854
5.5	0.0482	0.0484	-0.4%	0.0448	7.1%	0.0840
6.5	0.0489	0.0492	-0.6%	0.0461	5.7%	0.0814
7.5	0.0492	0.0497	-1.0%	0.0469	4.7%	0.0814
8.5	0.0489	0.0498	-1.8%	0.0469	4.1%	0.0837
9.5	0.0482	0.0491	-1.9%	0.0461	4.4%	0.0854
10.5	0.0480	0.0489	-1.9%	0.0458	4.6%	0.0850
11.5	0.0476	0.0485	-1.9%	0.0457	4.0%	0.0843
12.5	0.0479	0.0488	-1.9%	0.0458	4.4%	0.0873
13.5	0.0504	0.0513	-1.8%	0.0487	3.4%	0.0880
14.5	0.0519	0.0528	-1.7%	0.0503	3.1%	0.0913
15.5	0.0512	0.0520	-1.6%	0.0493	3.7%	0.0896

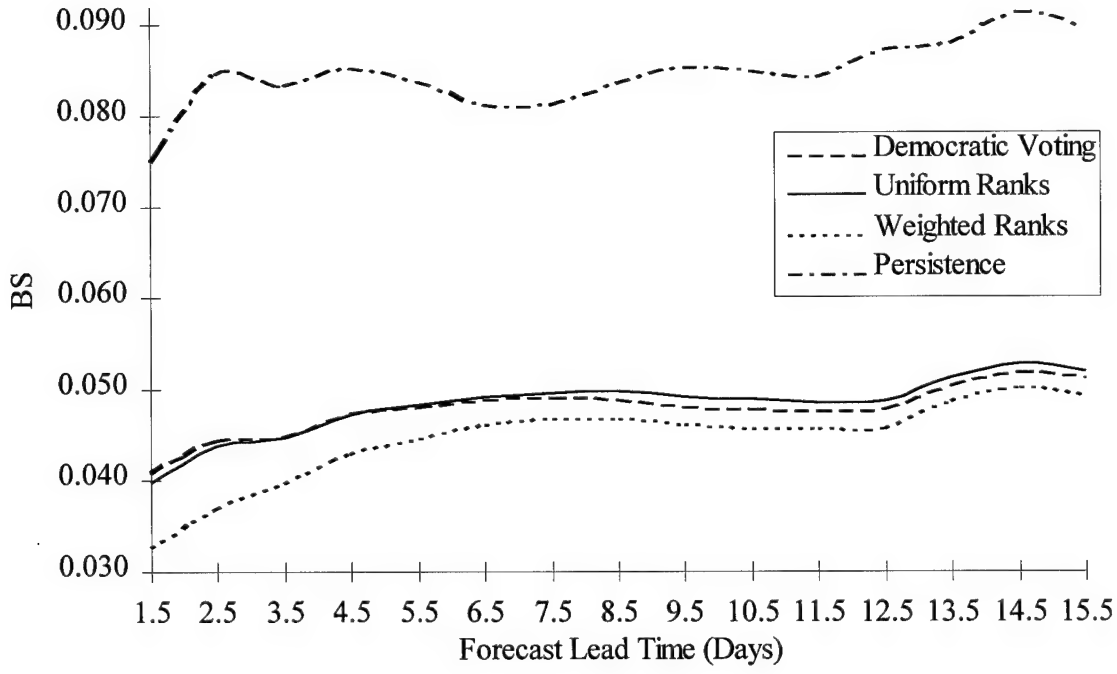


Figure 28. Graph of CAT3 Brier scores over the entire valid period.

Table 8. Brier scores for CAT4 for all lead times.

Lead Time (days)	Democratic Voting BS	Uniform Ranks BS / Improvement		Weighted Ranks BS / Improvement		Persistence BS
1.5	0.0101	0.0100	1.0%	0.0082	18.8%	0.0181
2.5	0.0105	0.0105	0.0%	0.0091	13.3%	0.0202
3.5	0.0099	0.0100	-1.0%	0.0091	8.1%	0.0202
4.5	0.0103	0.0104	-1.0%	0.0098	4.9%	0.0210
5.5	0.0109	0.0111	-1.8%	0.0104	4.6%	0.0213
6.5	0.0115	0.0117	-1.7%	0.0111	3.5%	0.0215
7.5	0.0111	0.0113	-1.8%	0.0108	2.7%	0.0206
8.5	0.0104	0.0107	-2.9%	0.0101	2.9%	0.0200
9.5	0.0100	0.0103	-3.0%	0.0097	3.0%	0.0203
10.5	0.0106	0.0109	-2.8%	0.0103	2.8%	0.0209
11.5	0.0111	0.0114	-2.7%	0.0108	2.7%	0.0218
12.5	0.0104	0.0108	-3.8%	0.0101	2.9%	0.0211
13.5	0.0118	0.0122	-3.4%	0.0116	1.7%	0.0225
14.5	0.0123	0.0126	-2.4%	0.0120	2.4%	0.0228
15.5	0.0118	0.0121	-2.5%	0.0116	1.7%	0.0225

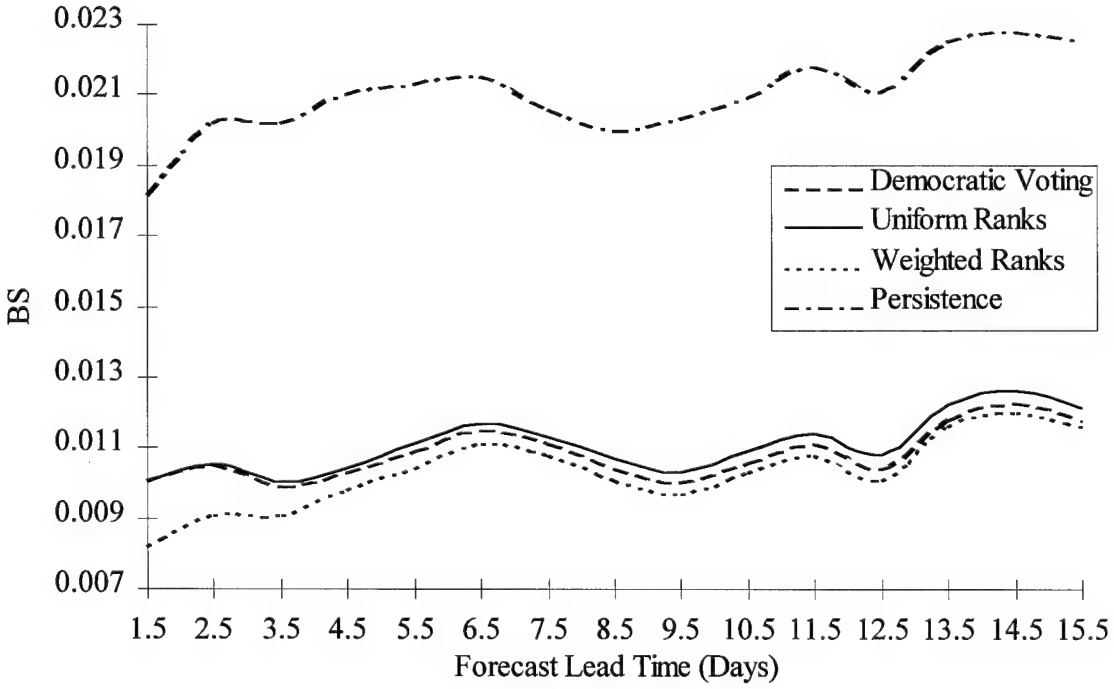


Figure 29. Graph of CAT4 Brier scores over the entire valid period.

that the BS is a mean square error measurement. For forecasting in CAT4, the sample space is dominated by cases where the verification did not occur above the threshold and the PQPF was 0.0%, resulting in 0.0 for a BS. These numerous low contributions cause an overall low mean BS. At lower thresholds, the proportion of these 0.0 contributions decreases so larger mean BSs result. In a sense, it is more difficult to forecast at the lower categories because the rate of occurrence is higher. This point is examined further in section c.

b) Ranked Probability Score

An overall score which is an extension of the BS to multicategory forecasts is given by the ranked probability score (RPS) (Wilks, 1995). It is a mean square error measure of probability forecasts which takes into account the error at each category. To do this, a probability forecast vector (PQPF at all 4 categories) is compared to an observation vector (0 or 1 for the same 4 categories) for each sample. For example, the forecast vector for the sample ensemble forecast *ENS* given in chapter 3 would be (0.80, 0.76, 0.49, 0.16). If the verification value were 15.1 mm, the observation vector would be (1.0, 1.0, 1.0, 0) since the verification occurred in CAT1, CAT2 and CAT3. The RPS is found by summing the squared errors of the components of the vectors for many samples as in Equation 9 (adapted from Equation 7.33b, Wilks, 1995).

$$\mathrm{RPS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{4} \left(FV_{i,j} - OV_{i,j} \right)^2 \tag{9}$$

where n is the total number of samples, i is an index for the sample number, j is an index for the category, FV is the forecast vector, and OV is the observation vector. For the

example, RPS = $(0.8 - 1.0)^2 + (0.76 - 1.0)^2 + (0.49 - 1.0)^2 + (0.16 - 0.0)^2 = 0.3833$. As with the BS, a perfect score is 0 while the worst possible score is 1.

The resulting RPS data for the research is presented in Table 9 and Figure 30 in a similar fashion to the BS results. Conclusions drawn from this data are the same as with the BS but worthwhile since the RPS gives an overall score rather than scores broken out by category. PQPF derived using the weighted ranks method is consistently the most accurate. The increase and convergence of the RPS of each method with increasing forecast lead time is even more evident than with the BSs.

An interesting side note concerns the plot of the persistence forecasts. First of all, notice the steep climb in the persistence RPS in the first 2.5 days of the valid period. This indicates that there were often occasions when the weather reoccurred for up to two days. In other words, the persistence forecast was occasionally correct. Further into the forecast valid period, there is a notable dip in RPS just past the 6.5-day lead time and again around 11.5-day lead time. This indicates that there is again a common reoccurrence of precipitation at these times in the valid period. This behavior of the persistence RPS is likely caused by the synoptic cyclone timescale (Hamill, 1998).

c) Reliability Diagram

A reliability diagram is a graphic display of the performance of a set of probabilistic forecasts for a certain category. Observed relative frequency of occurrence in the category is plotted as a function of forecast probability for the category (Wilks, 1995). For simplification and to increase the population at each probability bin, forecast probabilities are rounded to the nearest 10%. For an example plot value, consider 10

Table 9. Rank probability scores for all lead times.

Lead Time (days)	Democratic Voting RPS	Uniform Ranks RPS	Uniform Ranks Improvement	Weighted Ranks RPS	Weighted Ranks Improvement	Persistence RPS
1.5	0.2639	0.2563	2.9%	0.2247	14.9%	0.5123
2.5	0.2918	0.2858	2.1%	0.2563	12.2%	0.6154
3.5	0.3117	0.3067	1.6%	0.2842	8.8%	0.6242
4.5	0.3291	0.3252	1.2%	0.3089	6.1%	0.6296
5.5	0.3467	0.3438	0.8%	0.3319	4.3%	0.6257
6.5	0.3604	0.3578	0.7%	0.3470	3.7%	0.5998
7.5	0.3683	0.3663	0.5%	0.3578	2.9%	0.6029
8.5	0.3798	0.3785	0.3%	0.3714	2.2%	0.6349
9.5	0.3756	0.3749	0.2%	0.3694	1.7%	0.6412
10.5	0.3716	0.3712	0.1%	0.3633	2.2%	0.6418
11.5	0.3626	0.3626	0.0%	0.3585	1.1%	0.6161
12.5	0.3679	0.3680	0.0%	0.3659	0.5%	0.6346
13.5	0.3795	0.3792	0.1%	0.3787	0.2%	0.6441
14.5	0.3885	0.3882	0.1%	0.3866	0.5%	0.6657
15.5	0.3833	0.3828	0.1%	0.3805	0.7%	0.6533

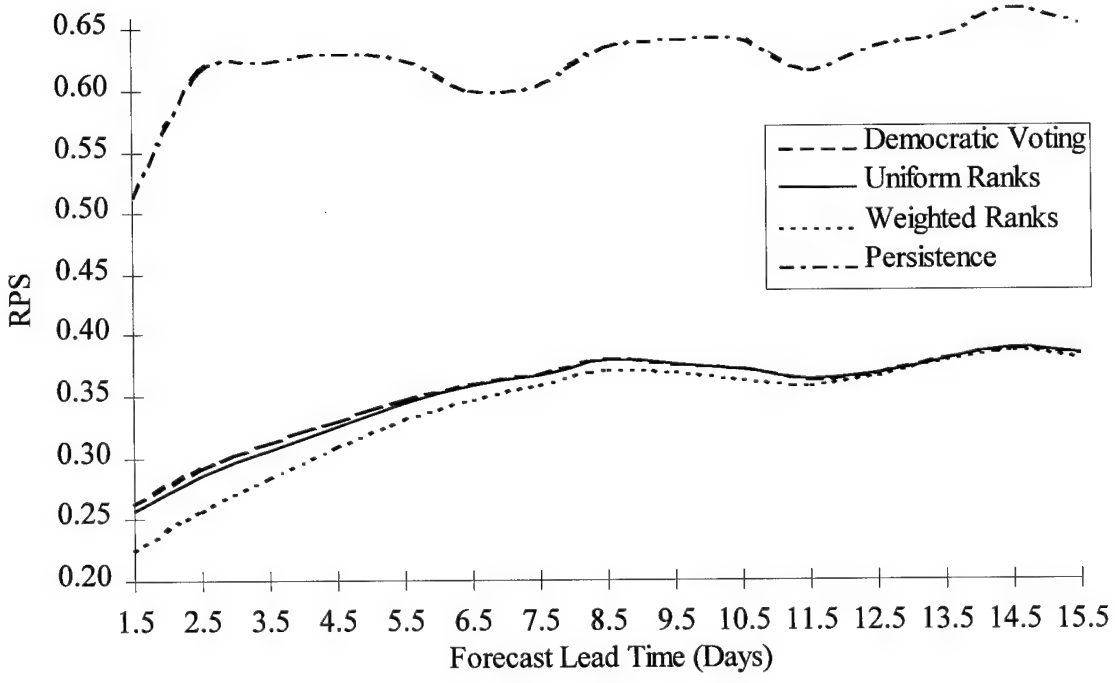


Figure 30. Graph of rank probability scores over the entire valid period.

different forecasts that all rounded to a 50% chance for $pcp24 > 2.54$ mm. For the set of these forecasts to be considered reliable, the observed relative frequency should be 5 out these 10 (50%) that verify with $pcp24 > 2.54$ mm.

Figure 31 gives an example of a reliability diagram from this research. For explanation purposes, this example displays much more information than is normally contained in the diagram. The plot of a perfectly reliable forecast is a line with a slope of 1.0 starting at the origin. The further the departure from this line, the less reliable the forecast. Points above the perfect line equate to underforecasting while points below equate to overforecasting. The histogram in Figure 31 (normally inset within the diagram) is a display of the relative frequency of usage of the forecast probabilities. It shows how many times 0%, 10%, 20%, etc. was forecast over the entire set of forecasts in the category.

Besides reliability, the diagram also displays the *resolution* and the *skill* of the forecasts (Wilks, 1995). These attributes are actually directly related to the BS. In fact, the reliability diagram is a visual break out of the BS which gives valuable information on the strengths and weaknesses of the forecasts. The BS can be decomposed into the three terms of reliability, resolution, and uncertainty as in Equation 10 (from Equation 7.28, Wilks, 1995).

$$\mathrm{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{11} N_i \left(FP_i - ORF_i\right)^2}_{\text{(reliability)}} - \underbrace{\frac{1}{n}\sum_{i=1}^{11} N_i \left(ORF_i - SC\right)^2}_{\text{(resolution)}} + \underbrace{SC\left(1 - SC\right)}_{\text{(uncertainty)}} \qquad (10)$$

where n is the total number of forecasts/observation samples for the category, i is the index for the 11 plotting points, N_i is the number of forecasts at each forecast probability,

Table 10. Raw data and reliability diagram data for CAT2, 1.5-day forecasts for PQPF derived from the democratic voting method. The observed (Obs.) relative frequency of occurrence is found by dividing the number of occurrences (Occ.) by the number of forecasts (Fcsts.). The percentage of occurrences is the number of occurrences divided by the total number of occurrences.

Forecast Probability	# of Fcsts	# of Occ.	Obs. Relative Frequency	% of Occ.	"Perfect Forecast" # 0f Occ.	% of Occ.
.0	15609	210	0.01	8.3	0	0.0
.1	1483	152	0.10	6.0	148	4.0
.2	884	121	0.14	4.8	177	4.8
.3	273	61	0.22	2.4	82	2.2
.4	457	102	0.22	4.0	183	5.0
.5	395	92	0.23	3.6	198	5.4
.6	369	130	0.35	5.1	221	6.0
.7	209	78	0.37	3.1	146	4.0
.8	595	267	0.45	10.5	476	12.9
.9	716	328	0.46	13.0	644	17.5
1.0	1412	990	0.70	39.1	1412	38.3
TOTAL:	**22402**	**2531**			**3687**	

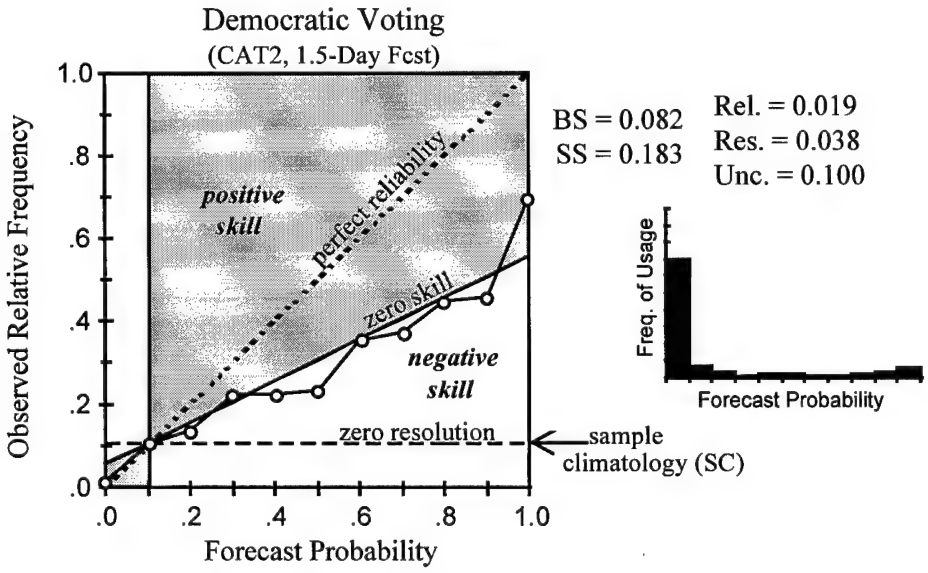


Figure 31. (adapted from Figure 7.8, Wilks, 1995) Reliability diagram for CAT2, 1.5-day PQPF derived from the democratic voting method. Dots (o) showing the observed relative frequency at each tenth of forecast probability are connected with line segments.

FP_i is the rounded forecast probability, ORF_i is the observed relative frequency, and SC is the sample climatology. This BS is only a close approximation to Equation 8 because rounded forecast probability is used. If rounded values were used in Equation 8, the results would be equivalent. This equation can best be understood by applying it to the data in Table 10 which made up the example diagram.

The *ORF* (plotted value on reliability diagram) is simply the number of occurrences at a particular *FP* divided by the number of forecasts, *N*. In the example, out of 884 20% forecasts there were 121 times that the observed value occurred in category 2. This gives an ORF = 121/884 ≈ 0.14, an overforecast. The reliability term is just a measure of the distance away from the perfect forecast line weighted by the number of forecasts at each *FP*. Better forecasts result in a smaller reliability term and thus a BS closer to zero.

The *SC* is the overall frequency of observed values occurring in the category. In the example, pcp_{24} occurred in category 2 a total of 2531 times out of the total of 22402 possible times. This gives an *SC* = 2531/22402 ≈ 0.11. In other words, this is the probability value a forecaster would give if he based his forecast solely on the climatological average occurrence. The resolution is a measure of distance away from the climatological probability forecast (dashed line labeled zero resolution) weighted by the number of forecasts at each *FP*. Better forecasts result in a larger resolution term and thus a BS closer to zero.

The uncertainty term (plotted in Figure 32) is set by the *SC* and thus independent of the forecast. It can be thought of as a measure of how easy it is to forecast the event.

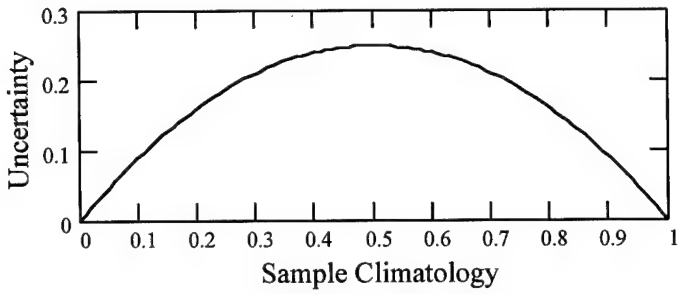


Figure 32. Graph of the uncertainty term of the BS. Maximum uncertainty of 0.25 occurs at a sample climatology of 0.5.

The highest uncertainty (most difficult to forecast) is an event that occurs half of the time on average. An event that rarely occurs or frequently occurs has a lower uncertainty (easier to forecast). The extreme is an event that either never occurs or always occurs which is of course quite simple to forecast for and thus has an uncertainty of zero.

The skill score (SS) of the forecasts is determined by (resolution – reliability) / uncertainty (Equation 7.29, Wilks, 1995). Therefore, for a point to contribute positive skill, it must have a reliability value smaller than its resolution. This is represented by the shaded region in Figure 31. Outside of this region, reliability is larger than resolution and skill is negative. Forecasts that exhibit an overall negative SS performed worse than a climatological forecast. In other words, PQPF based simply on the climatological norms would be a higher quality product.

Figure 31 is an example of forecasts which are somewhat reliable but have rather low skill. It is clear that even for forecasts with only a 1.5-day lead time, the uncalibrated democratic voting method produces low quality PQPF. The information in Figure 31 is repeated in the more standard reliability diagram format in Figure 33b. The arrows next

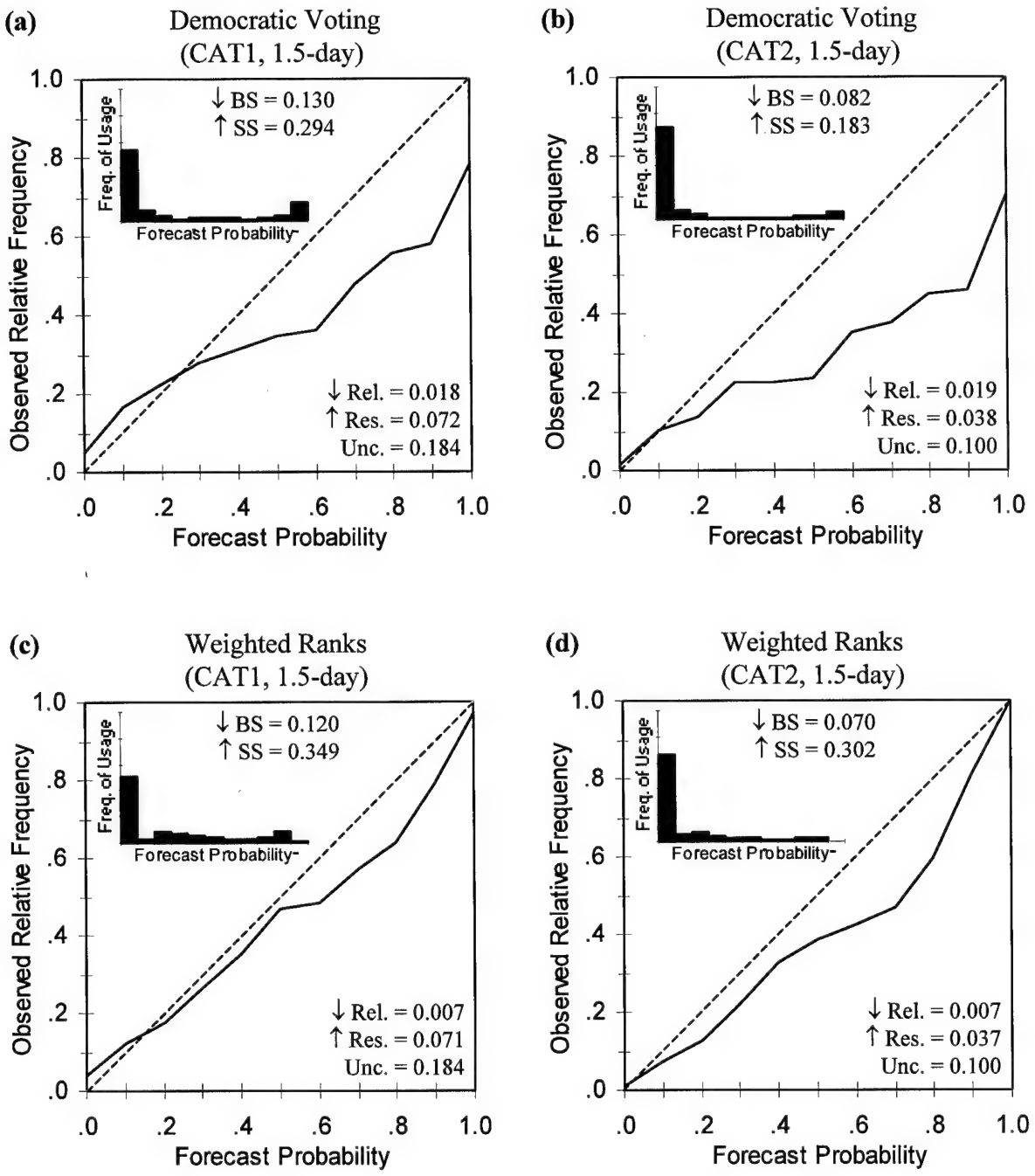


Figure 33. Reliability diagrams for all 4 categories of 1.5-day forecasts for uncalibrated (democratic voting) and calibrated (weighted ranks) PQPF. Panels are organized for vertical comparison (E.g.: CAT1 for democratic voting (a) is directly above weighted ranks (c)). Arrows before score names remind the reader of the desired direction for better forecasts. Up arrows (↑) where a higher number is better and down arrows (↓) where lower is better.

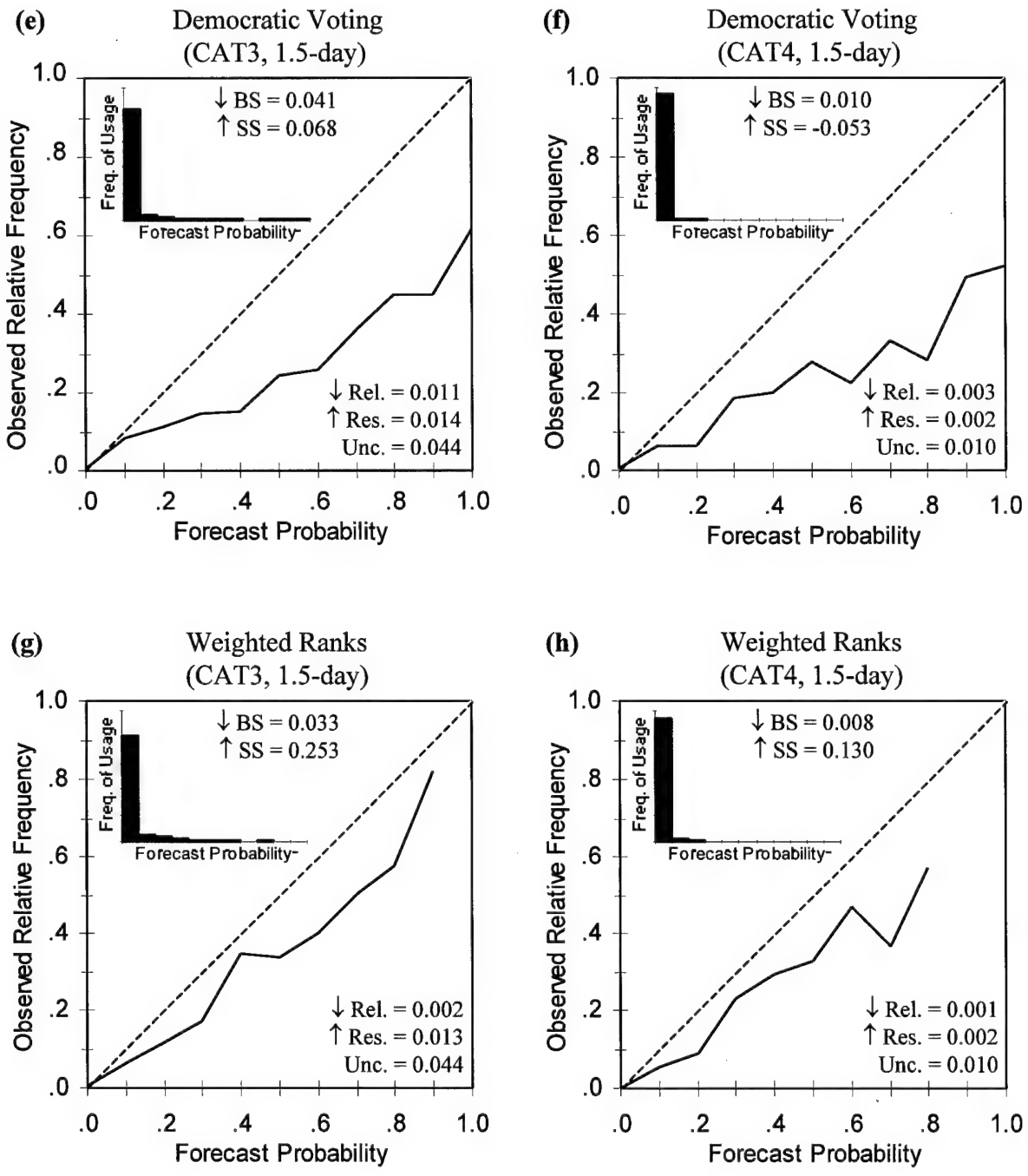


Figure 33. (continued)

to the BS, SS, reliability (rel.), and resolution (res.) are reminders of the desired quantitative direction for higher quality forecasts.

Reliability diagrams for all four PQPF categories for 1.5-day forecasts are displayed in Figure 33 and for 3.5-day forecasts in Figure 34. Only the democratic voting and weighted ranks PQPF results are shown to reduce the clutter. Also, the relative quality of the uniform ranks and persistence PQPF has already been well established with the BS and RPS. The lead time of 1.5 days was selected to be displayed since 1.5-day forecasts showed the most significant improvements. The lead time of 3.5 days was selected since it is in the medium range and shows several interesting features.

The most important feature of these diagrams is that the calibrated PQPF of weighted ranks method is shown to consistently outperform the uncalibrated PQPF of the democratic voting method. This is evident in the line plots of observed relative frequency as well as with the BSs and the SSs. However, it is also evident that the calibrated forecasts exhibit the same basic problem as the uncalibrated forecasts, a tendency towards overforecasting of probability.

Another important point to be drawn from these diagrams concerns what happens to the quality of PQPF as the category threshold increases. In the previous section, it was noted that the BS is lower for the higher category thresholds which appears to indicate that PQPF is better at higher thresholds. The reason for the lower BS at higher thresholds is simply because of the lower uncertainty. The BS is only useful for comparing different forecasts at one threshold, not comparing different thresholds of the same forecasts.

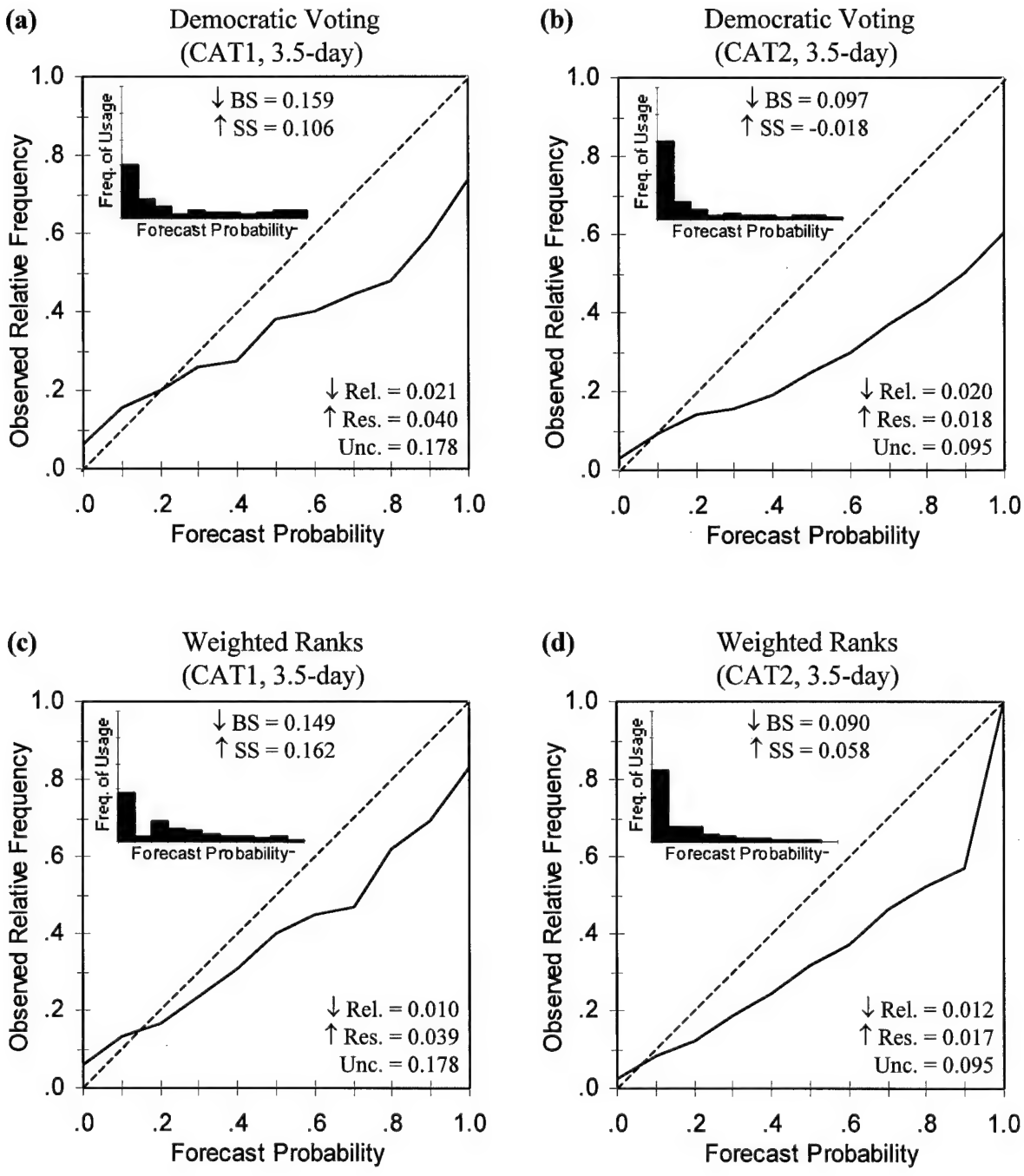


Figure 34. Reliability diagrams for all 4 categories of 3.5-day forecasts for uncalibrated (democratic voting) and calibrated (weighted ranks) PQPF.

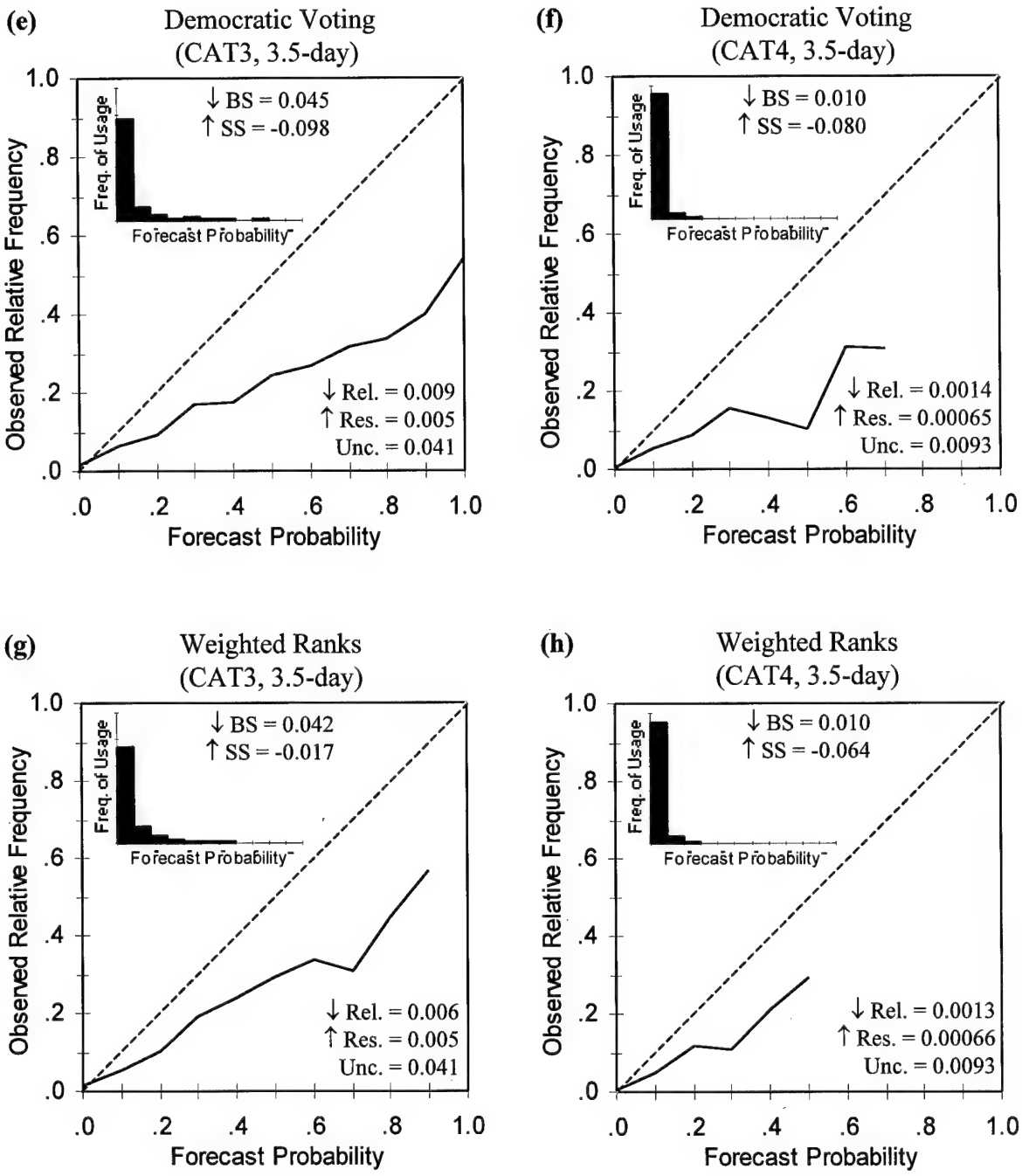


Figure 34. (continued)

The SS is useful for comparing both different forecasts and different thresholds. From the trend in SS and the reliability diagram plots, it is now clear that the quality of PQPF decreases as the category threshold increases. To examine these results in more detail, Figure 35 plots the SS of both PQPF methods for the four categories over the entire valid period.

These plots show that the calibration did do its job (improved skill of PQPF) but only to a limit. Recall that a negative SS indicates forecasts with predictive skill below that of forecasts based on climatology. In the CAT1 SS plot (Figure 35a), both PQPF SSs drop below the zero line soon after a lead time of 5.5 days. Beyond this point then, neither forecast is of any use since a climatological forecast is more skilled. The calibration's ability to improve PQPF appears to be limited by the general predictability of cumulative precipitation. The calibration can extend this predictability to a certain degree. A more dramatic extension of predictability by the calibrated PQPF is seen in the higher thresholds. The usefulness of PQPF is advanced by over one day in CAT2 and CAT3. In CAT4, the uncalibrated forecasts were totally unskilled while the calibration managed to briefly exhibit positive skill.

d) Confidence Diagram

There is a very important aspect of probability forecasts that the reliability diagram does not directly display. That is, of the times an event actually occurred, how much probability was typically given for the chance of occurrence? The reliability diagram gives results for all forecasts, whether the event occurred or not, thus missing

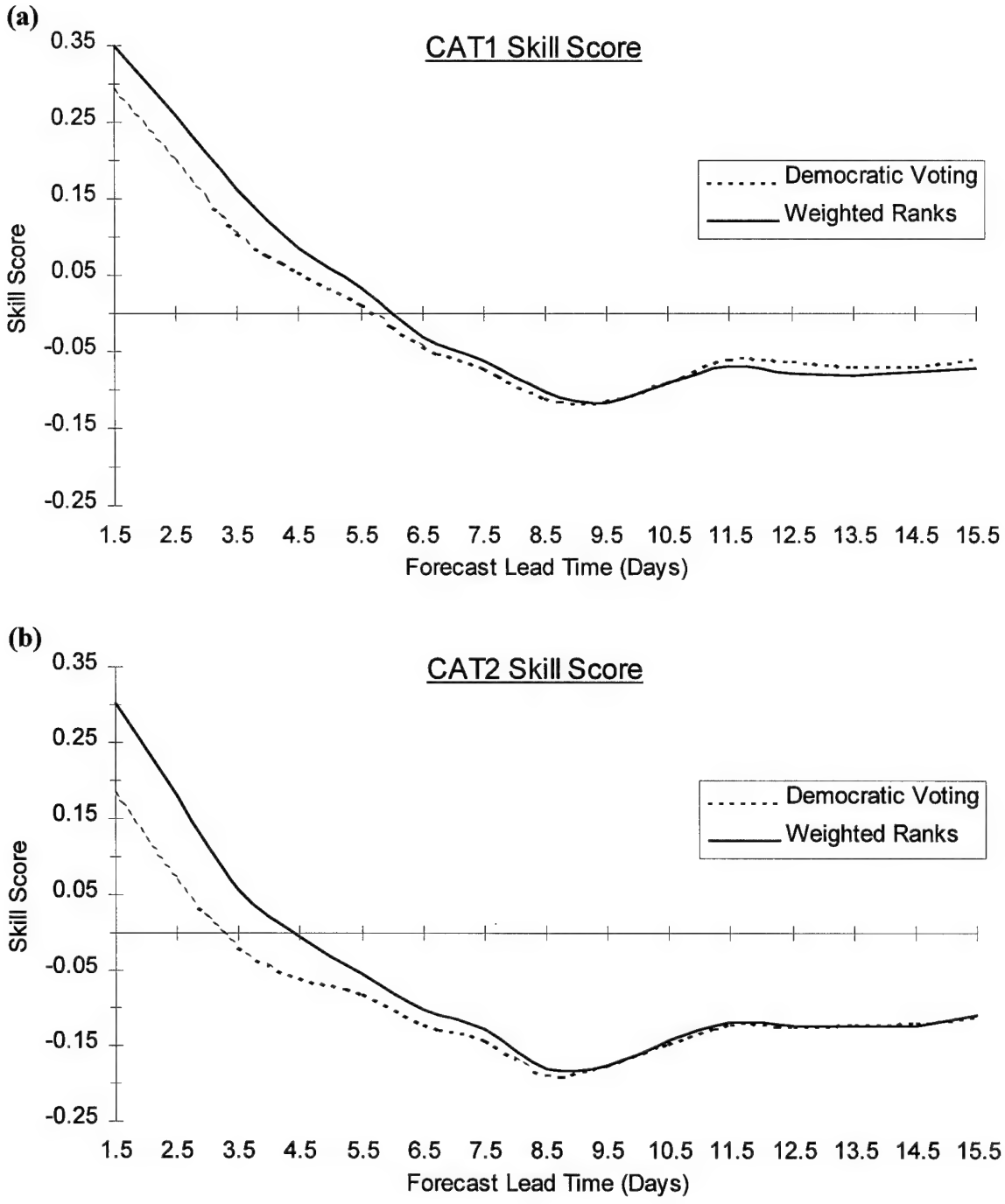


Figure 35. Skill scores over the entire forecast valid period determined from reliability diagrams of uncalibrated (democratic voting) and calibrated (weighted ranks) PQPF. A score below zero indicates forecasts that have a predictive skill below a forecast based on climatology.

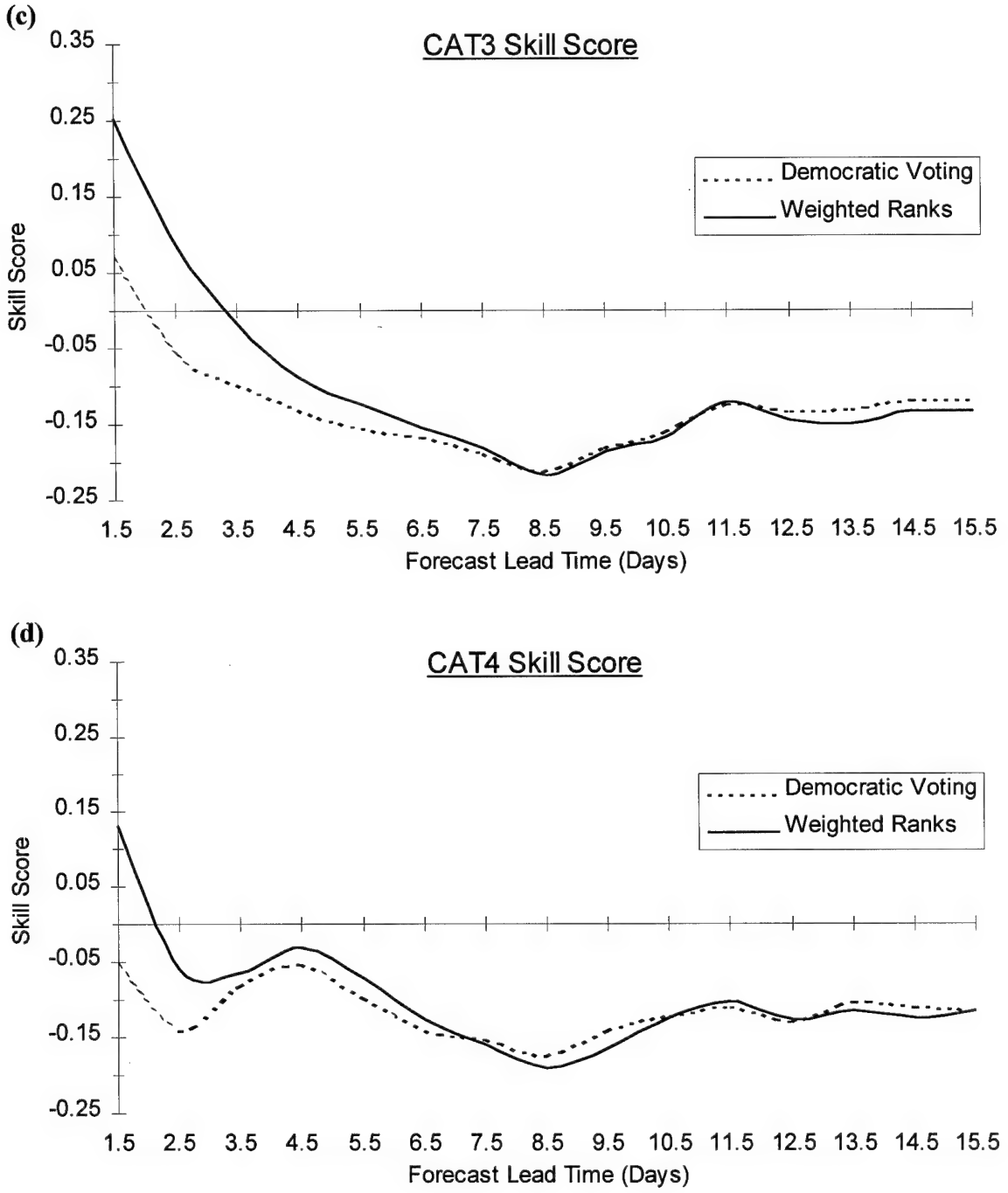


Figure 35. (continued)

this point. While the BS and the SS do reflect this aspect of the forecasts, a visualization of this factor is necessary to comprehend its significance.

It is possible to make highly reliable, positively skilled forecasts that are not very useful from the customers point of view. This is because in general, weather forecasters and their customers are much more concerned with the occurrence of an event as opposed to the non-occurrence. For this reason, a measurement tool for probabilistic forecasts was designed called the confidence diagram. The name comes from the fact that it shows to what degree of confidence the forecasts were made. It also relates how much confidence a customer would put in a forecaster's ability to forecast the event.

Before discussing the new diagram, the need for the diagram must be clearly explained. Consider a hypothetical example of two different forecasters, *A* and *B*, who made predictions on the chance of the temperature rising above 38°C in Atlanta during 5 days in August. Forecaster *A* made 5 forecasts all at 20% chance. Forecaster *B* made 4 forecasts at 0% and the last one at 100% chance. The observation was that days 1 – 4 were below 38°C and day 5 was above. Both forecasters achieved perfect reliability. However, forecaster *B* achieved a better skill score because of a high resolution. Obviously, the customer would be more happy with *B*'s forecasts than *A*'s. With this simple example, it is easy to comprehend the difference in the forecasts. With large samples of forecasts which use all the probability values, it gets difficult to understand the meaning behind the skill score.

The confidence diagram is a histogram of how the occurrences (events where *pcp24* occurred in the category) were forecast. In a large sample of probabilistic

forecasts, the event will likely occur for all forecast probability bins. If these forecasts are highly reliable, the observed relative frequency of the occurrences will be close to forecast probability value in each bin. What is desirable is to have most of the occurrences in the bins of higher levels of forecast probability. In the above example, forecaster *A* had a low confidence so the 1 occurrence was forecast at 20%. Forecaster *B*, with high confidence, correctly forecast 1 occurrence with a 100% forecast.

Referring back to the sample data in Table 10 on page 86, the percent of occurrences column was computed by dividing the number of occurrences at the forecast probability by the total number of occurrences. This information is shown in the black histogram bars of Figure 36b. In this case, the majority of the occurrences were in the higher forecast probabilities, the desired characteristic. However, a significant portion of the occurrences was forecast at lower percentages. Also shown are the results for the same forecasts if they had been perfectly reliable. The gray shaded histograms show percentage of occurrences for a set of perfectly reliable forecasts that have the same numbers of forecasts at each forecast probability as the actual forecasts. These perfectly reliable forecasts are not the set of forecasts with both perfect reliability and perfect accuracy (i.e., forecast set containing only 0% or 100% forecasts with perfect reliability).

Figure 36 and Figure 37 display confidence diagrams for the same categories and lead times as in Figure 33 and Figure 34. The reliability and the SS are repeated since their values are visually depicted in the diagrams. Both the uncalibrated and calibrated PQPF show a steady decrease in confidence with increased threshold and increased valid time. This is evident in the steady shifting away from the higher forecast probabilities.

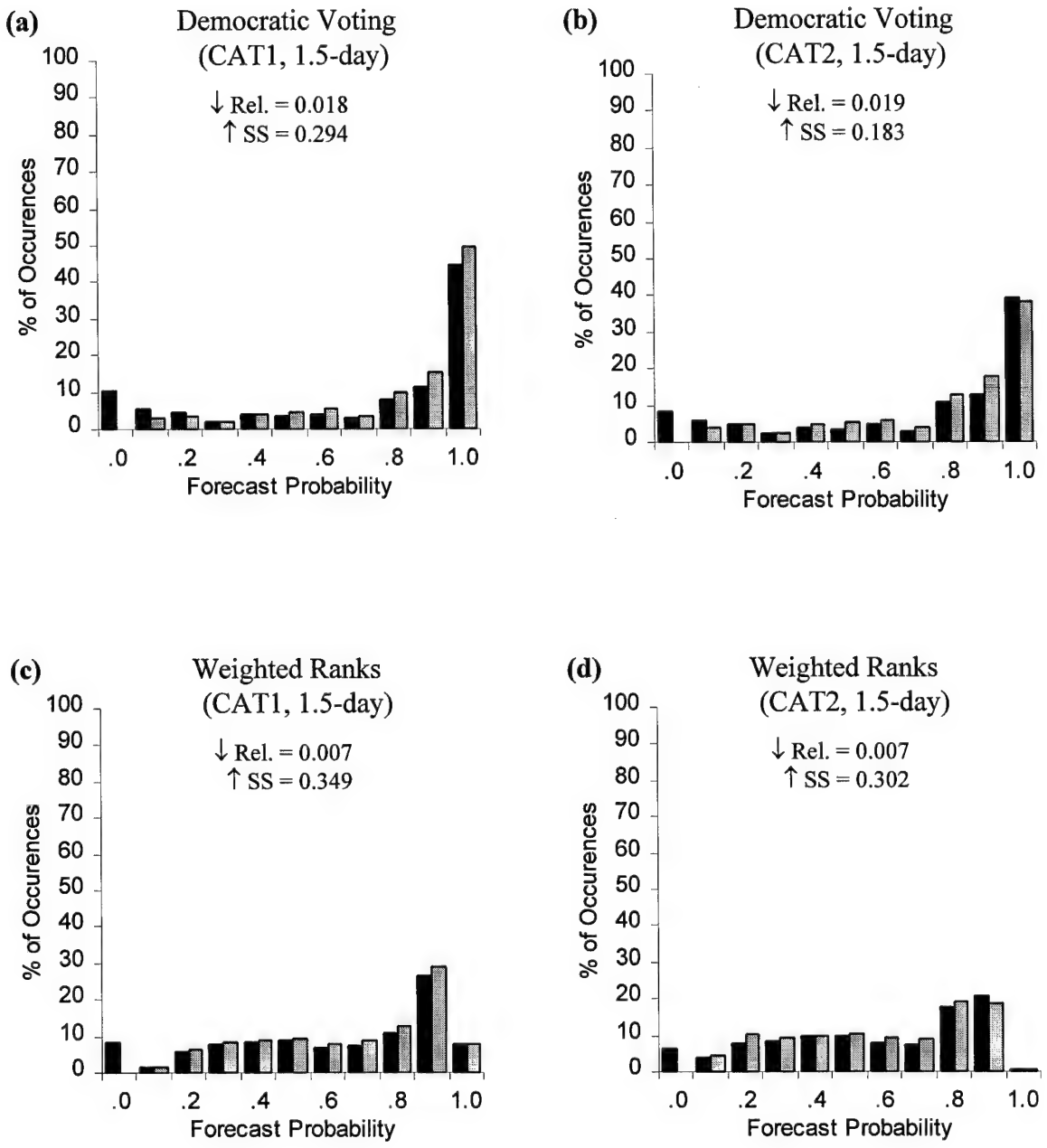


Figure 36. Confidence diagrams for all 4 categories of 1.5-day forecasts for uncalibrated (democratic voting) and calibrated (weighted ranks) PQPF. Black bars are percentages of the occurrences for the actual set of forecasts while gray bars are for the same forecasts, but with perfect reliability.

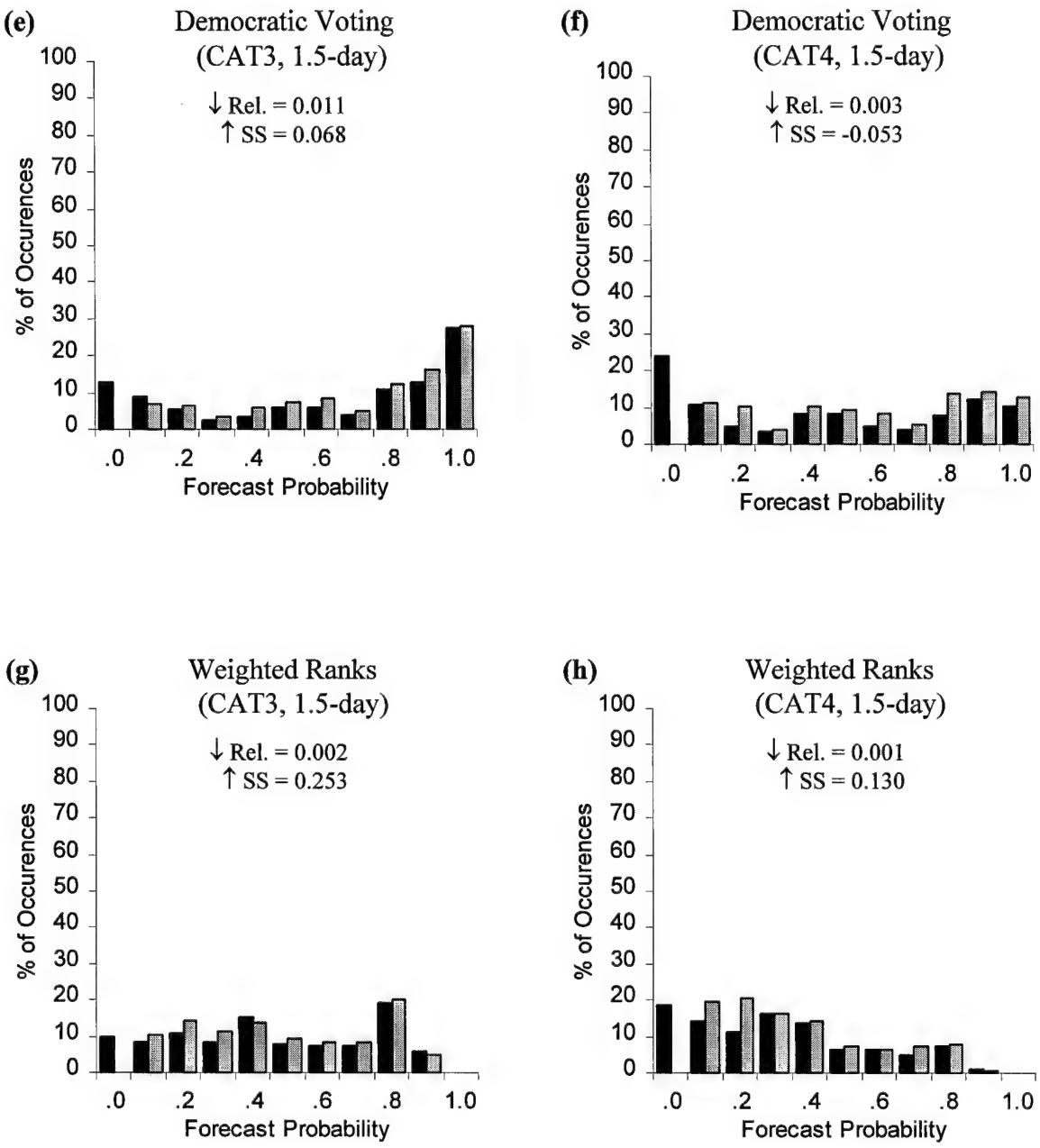


Figure 36. (continued)

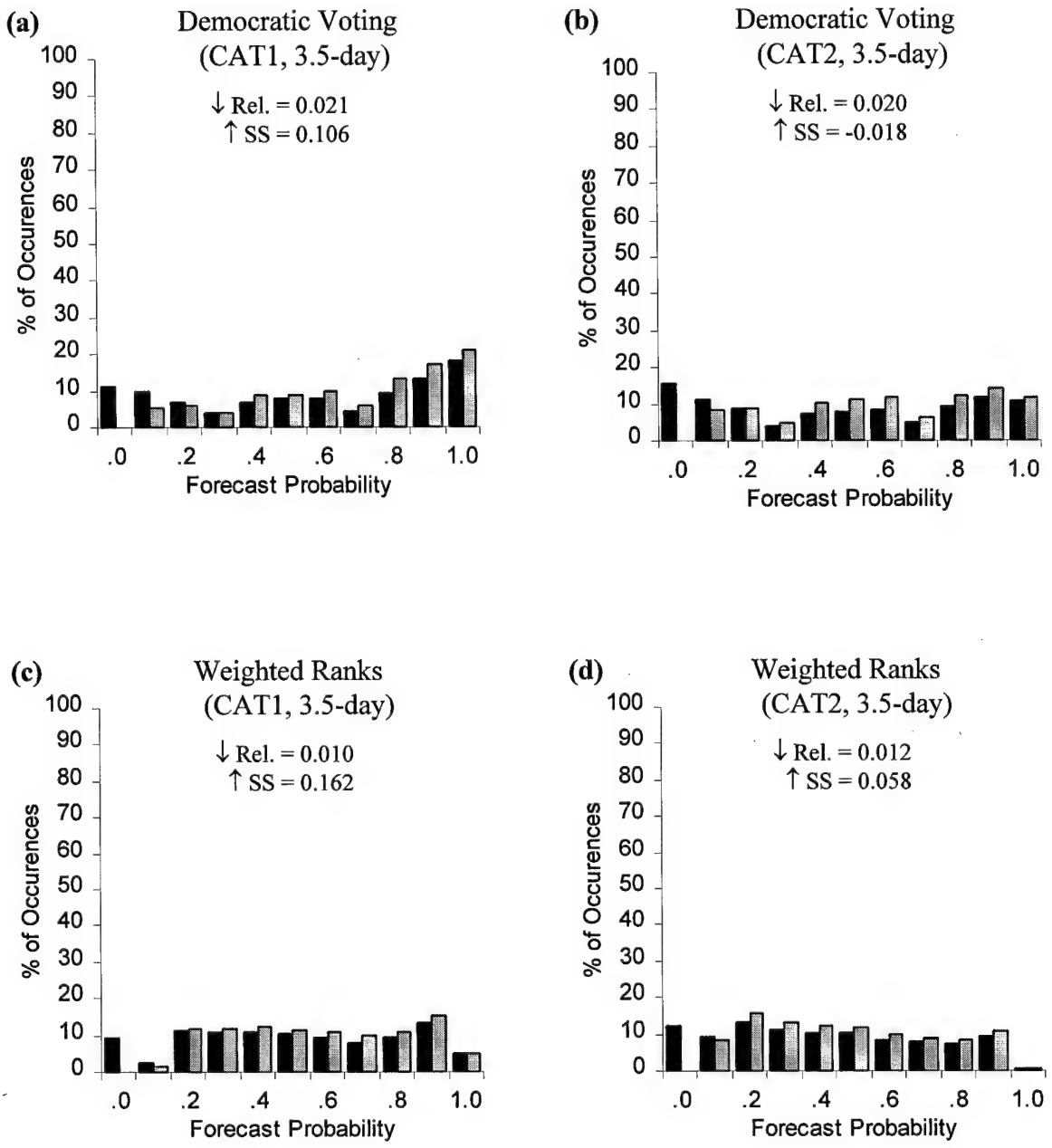


Figure 37. Confidence diagrams for all 4 categories of 3.5-day forecasts for uncalibrated (democratic voting) and calibrated (weighted ranks) PQPF.

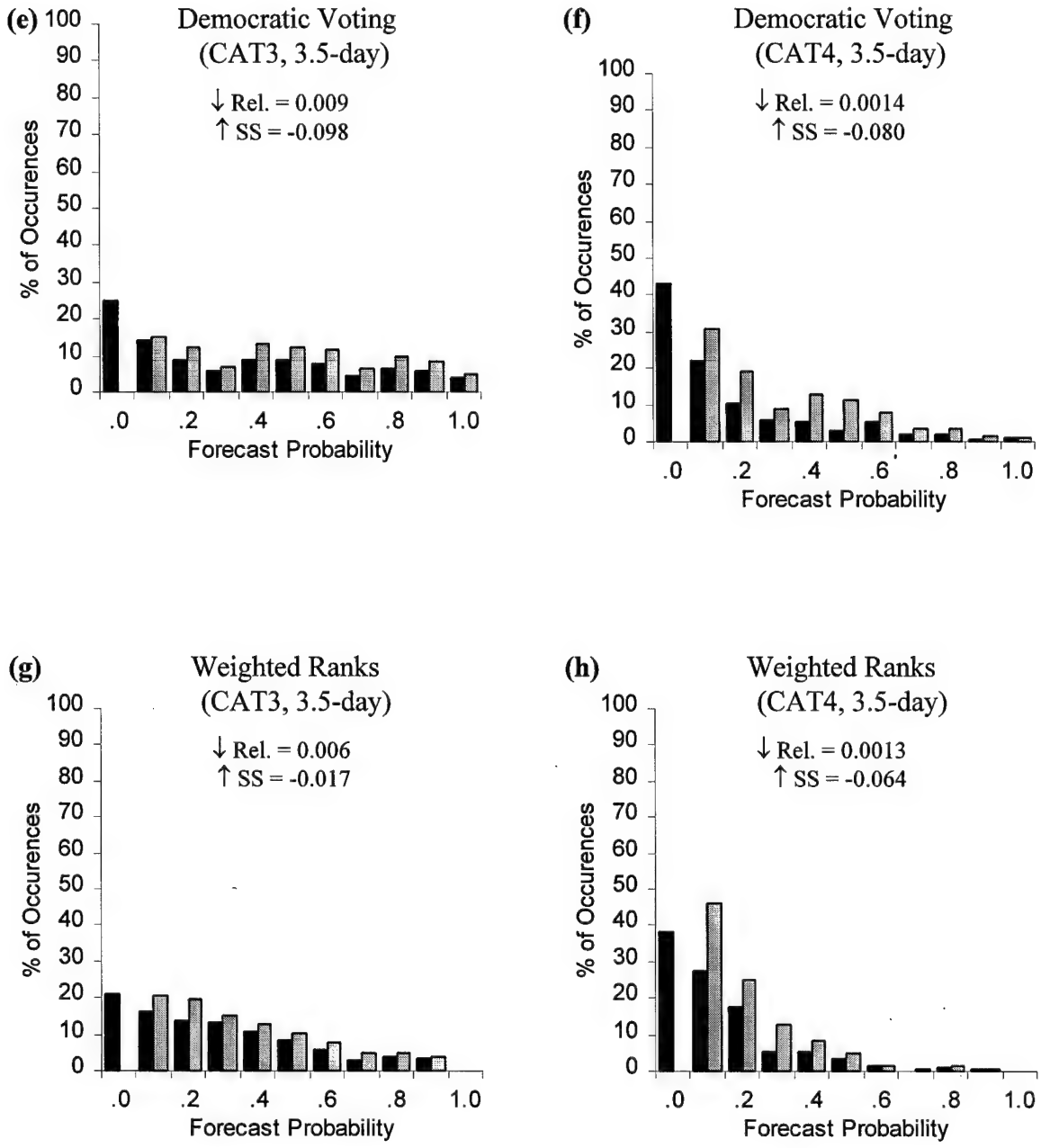


Figure 37. (continued)

Consider the democratic voting CAT4, 1.5-day forecasts (Figure 36f). What makes these forecasts so unskilled is that in nearly 1 out of 4 times that *pcp24* measured greater than 25.4 mm, the PQPF was 0%. This is clearly not a useful forecast product. While the calibration (Figure 36h) improved upon the percentage of occurrences at 0% forecast probability, its overall usefulness is also questionable since the majority of occurrences were given 50% chance or less.

An additional aspect of these diagrams is that they reveal the effect of the calibration on the PQPF. Compare any set of uncalibrated and calibrated confidence diagrams. Notice that the calibration consistently shifts the percentage of occurrences from the extremes toward the center. This means that the calibration decreases the confidence for the higher PQPF percentages and increases it for the lower PQPF percentages. In this way, an overall higher quality PQPF is created.

d. Limits of Predictability

Referring back to Figure 35 on page 95, the limits for skillfully predicting cumulative precipitation with the MRF ensemble are approximately 6.0 days for CAT1, 4.4 days for CAT2, 3.3 days for CAT3, and 2.1 days for CAT4. This finding brings up two questions: (1) What is the relationship between *pcp24* threshold and predictability; and (2) is this finding simply due to the limitations of the MRF ensemble or is it a more fundamental truth?

Figure 38 was designed to help answer the first question. After a log transformation of the *pcp24* threshold values, a third-order polynomial function was fit to the four limits of predictability described above. While the curve is an excellent fit, it

should be taken only as a gross approximation in the extrapolated regions to the left and right of the four data points. The curve indicates that predictability falls off sharply at first with increasing *pcp24* threshold then decreases more gradually at higher levels of *pcp24* threshold.

This result is consistent with the ideas presented by Lorenz (1969) who proposed that smaller scales of motion have shorter ranges of predictability. Low levels of *pcp24* are generally associated with widespread precipitation events occurring on the synoptic scale. An event with a large *pcp24* amount occurs on a smaller scale, more likely due to convective activity than a synoptic scale storm. Therefore, predictability should decrease with increasing *pcp24* threshold (i.e., decreasing scale) as displayed in Figure 38.

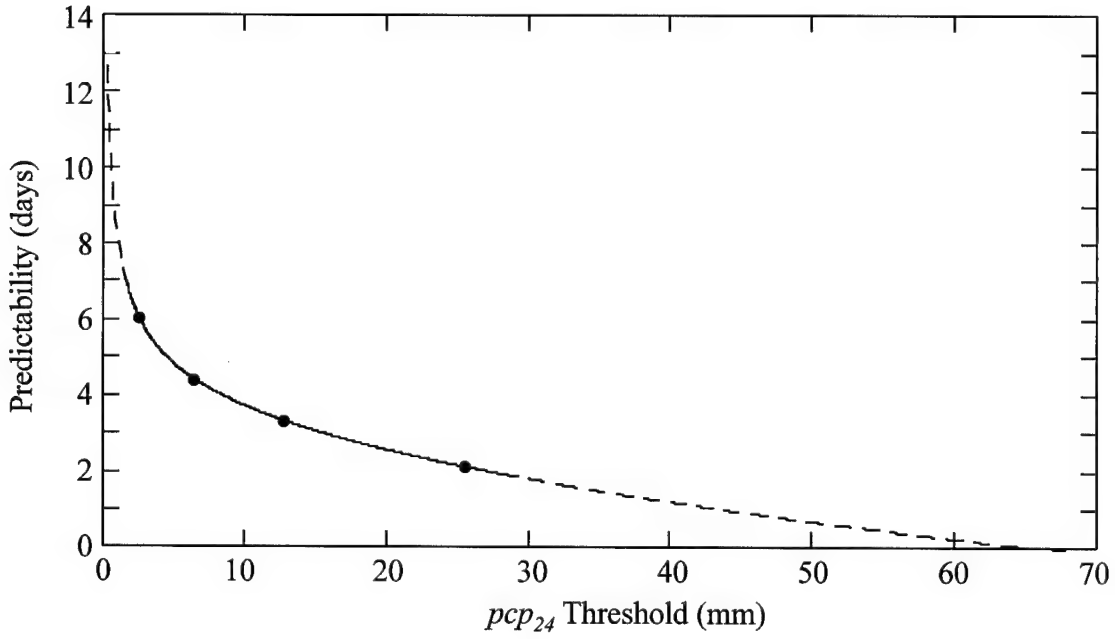


Figure 38. Limit of predictability as a function of *pcp24* threshold. Results from skill score analysis are plotted as dots (•). The solid line part of the curve shows the more credible results while the dashed parts of the curve are less credible extrapolations.

Lorenz (1969) also described the limits of predictability as intrinsic to a chaotic dynamical system, and can not be increased by reducing observational error. This helps in answering the second question which can now be rephrased as: Are the limits of predictability displayed in Figure 38 the intrinsic limits which exist because the atmosphere is a chaotic dynamical system? It is logical to presume that the limits analyzed here are shorter than the true limits due to deficiencies in the MRF ensemble.

While no definitive answer can be provided within the context of this research, a reasonable supposition can be made. It is true that the MRF is not perfect and the ensemble IC are less than ideal, but it is also true that the calibration was designed to compensate for these inadequacies. That was in fact the whole point of the calibration. Therefore, the limits of cumulative precipitation predictability indicated by the analysis of calibrated PQPF skill score may be close to the true limits.

It should be noted that the limits of predictability shown in Figure 38, whether they are the true limits or not, represent the average predictability over a large sample space. The limits can vary depending upon the local divergent nature of the atmosphere's attractor for a particular forecast period. In other words, some weather patterns are much more predictable than others even though they are of similar scale.

e. Difficulty of Probabilistic Forecasting

In the discussion of the decomposition of the BS earlier in this chapter, the idea was introduced that uncertainty in the occurrence of an event contributes to the difficulty in forecasting that event. Referring back to Figure 32 on page 88, events which rarely or

always occur were described as easier to forecast since a forecaster and/or model would generally be more certain of the future. However, this is not the whole story.

Consider the confidence diagrams displayed in the previous section of this chapter. If CAT4 *pcp24*, an event of rare occurrence, is easy to forecast for, why is the confidence and the skill of these forecasts so low compared to the other categories at the same lead time? The answer is that the difficulty in making a probabilistic forecast must be a combination of uncertainty and *targetability*. This term is so named because forecasting an event can be considered like trying to hit a target. A big target (an event that often occurs) is easy to hit. As the target gets smaller (event occurs less often) or further away (increased lead time), it quickly gets harder to hit the target.

The term for targetability is therefore a function of both the sample climatology (*SC*), or average percent occurrence of the event, and lead time (*t*). Compared to the uncertainty term which has quantitative meaning, targetability is only a qualitative term based on logic and experience of the author. The qualitative function for forecast difficulty (Equation 11) is simply a linear combination of the two terms.

$$\text{Forecast Difficulty} = \underbrace{t\left(\frac{1-SC}{SC}\right)}_{\text{(targetability)}} + \underbrace{SC(1-SC)}_{\text{(uncertainty)}} \qquad (11)$$

The plot of the forecast difficulty function at one lead time (Figure 39a) explains why the quality of the PQPF gets worse for higher *pcp24* thresholds. The targetability term behaves asymptotically so its influence is small for common events (high *SC*) but dominates for rare events (low *SC*). PQPF for CAT4 are made with low uncertainty but they are extremely hard to target since CAT4 rarely occurs. This makes CAT4 much

more difficult to forecast. Furthermore, targetability gets harder with increased lead time as shown in Figure 39b, making forecasting at any category more difficult.

While there is no hard scientific proof of the validity of the forecast difficulty function, it does offer an intuitive understanding for the challenges involved in probabilistic forecasting. Additionally, it provides another explanation for the trends in PQPF quality analyzed in section c of this chapter as well as the findings concerning the limits of predictability.

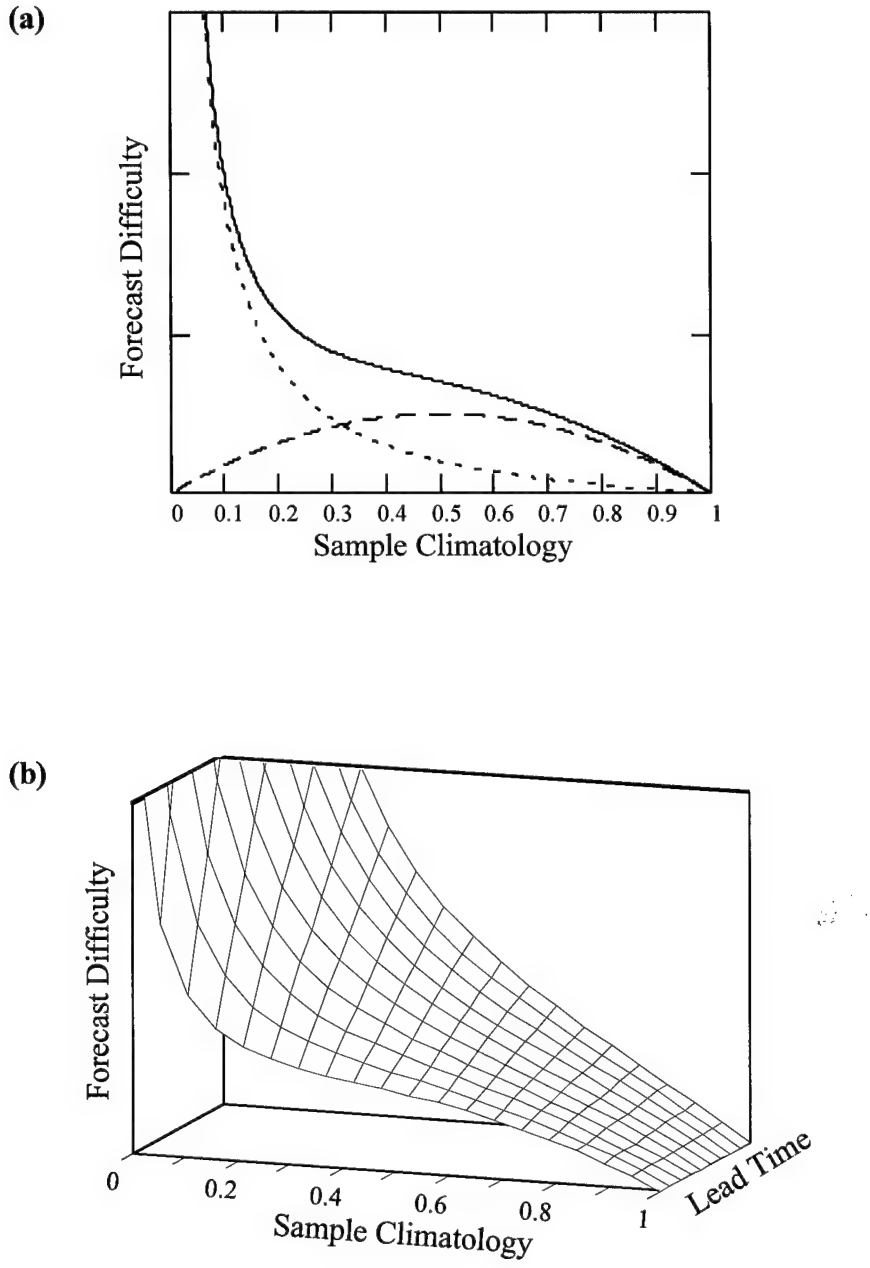


Figure 39. Plot of the forecast difficulty function. Forecast difficulty and lead time axes are not numbered because this plot is qualitative. (a) Forecast difficulty (solid curve) is a combination of the uncertainty term (long dashed curve) and the targetability term (short dashed curve). (b) Forecast difficulty increases with lead time due to time dependence of targetability term.

5. Conclusions and Recommendations

a. Overview

This chapter begins by presenting the major conclusions drawn from the results of this research. It goes on to discuss recommendations concerning the specific subject of this thesis as well as ensemble forecasting in general. Lastly, three suggestions for future research are given.

b. Conclusions

Calibrated PQPF produced by the weighted ranks method does dramatically improve the quality of PQPF, as was clearly shown by the findings of the various measurement tools discussed in chapter 4. Unfortunately, the calibration can only do so much. In comparing the calibrated and uncalibrated PQPF, it was discovered that the limit of predictability for significant amounts of cumulative precipitation is much shorter than was originally surmised.

Referring back to Figure 38 on page 104, it was found that for higher levels of cumulative precipitation, of more concern to meteorologists and their customers, it is not possible to produce reliable and/or accurate PQPF based on the MRF ensemble in the medium range. The spatial and temporal variability of high amounts of precipitation is too great to be forecast with skill beyond the short range. However, PQPF based on the MRF ensemble is of value in the medium range for very low thresholds of cumulative precipitation. In other words, in the medium range it is possible to forecast the probability of precipitation vs. dry conditions with skill but not possible to predict

precipitation quantity. These conclusions may apply to all forecasts of cumulative precipitation, not just to forecasts based on the MRF ensemble.

c. Recommendations

Concerning the focus of this research, there are two recommendations: (1) PQPF based on the MRF ensemble should not be employed in the medium range for any purpose other than the lowest possible threshold PQPF (i.e., cumulative precipitation > 0.1 mm); and (2) the calibration designed by this research should be implemented by NCEP to maximize the skill of PQPF over the United States. This is of course provided that the calibration designed in this thesis is more robust than the technique developed at NCEP. A comparison of the skill found in this research to the skill of NCEP's calibrated PQPF needs to be accomplished to answer this question. Whichever technique proves better, that calibration process should then be applied to other regions of the world where improved PQPF would be beneficial.

A more far reaching recommendation concerns the overall science of ensemble forecasting and the USAF: The Air Force Weather Agency should seek to employ ensemble forecasting in all aspects of weather operations. Instead of focusing on improving NWP with smaller scale models, more effort must be made in researching how the USAF can benefit from ensemble forecasting. The fact that short range ensemble forecasting is still in development is even more reason to get involved now.

The most obvious benefit of using ensemble forecasting would be that forecasters would have a better comprehension of the confidence in a particular model, which should be known when using any NWP product. Determination of confidence in a model is

currently done to some extent by USAF forecasters by comparing different models or previous runs of the same model and by examining a model's performance early in the forecast period. Such techniques are worthwhile but quite inferior compared to the information provided by spaghetti diagrams or standard deviation charts from an ensemble forecast.

Other potential benefits of short range ensemble forecasting are both promising and exciting. Using an ensemble mean as the best forecast over the control represents a way to improve a model's output without increasing its resolution. Probabilistic forecasts based on an ensemble (as presented in this research) for any weather parameter are a new concept for USAF weather operations and have countless applications.

d. Future Research

Since ensemble forecasting is still a new application of NWP, there is a tremendous amount of research yet to be done before the full potential of the technique can be realized. Concerning the narrow sub topic of this thesis (calibration of probabilistic forecasts based on an ensemble), there are three areas for possible future research.

(1) If a more effective calibration could be designed, the quality of PQPF would be improved even further. Such a calibration may be similar to the one presented in this thesis but based on more than just ensemble standard deviation. In general, the more flexible and detailed the calibration, the more accurate and reliable the probability forecast. Of course, this improved quality comes at a price. To design a better calibration, a bigger sample space is required along with a more detailed analysis,

requiring more effort. A trade-off between improved quality and design effort would therefore need to be made. The most straightforward choice would be to design a calibration based on weather regimes as discussed in chapter 4.

(2) If it could be confirmed that the weighted ranks calibration technique works well for probabilistic forecasts of different weather parameters (i.e., temperature, wind speed, visibility, etc.), then ensemble based probabilistic forecasts would have many more applications. There is no reason why such forecasts should be restricted to precipitation. For instance, say Kadena AFB, Japan, is concerned about the possibility of damaging winds from a typhoon that may pass nearby in several days. A probabilistic forecast for winds > 50 kt over a particular time period could be based on a ensemble forecast. This ensemble would of course contain systematic errors which could be effectively accounted for by a calibration technique just like the one presented in this thesis. Theoretically, the result would be a more accurate and reliable probability forecast for damaging winds and consequently, more effective resource protection actions could be taken.

(3) If the methods of this research were repeated with better observational data, the findings would be more credible. Once the problems with the cumulative precipitation data from the multisensor analysis are solved, it would be the best choice to represent the true precipitation. A new study using this data would have the same general conclusions as this thesis but would be able to give more detailed results concerning the true limits of predictability of cumulative precipitation.

Appendix A: Polynomial Coefficients of Probability Surfaces

This appendix lists the coefficients of the third-order polynomial curves which were fit to the 18 ranks at each of the 15 lead times in the forecast valid period. The coefficients were generated with a Mathcad template then saved to a data file for use by the Fortran program PQPF (see Appendix B). The coefficients were taken to ten decimal places since regular precision (seven digits) was used in the Fortran program. Most of the smallest coefficients begin in the $10000^{\underline{ths}}$ place thus yielding seven digits. Larger coefficients carry more precision than could be used by the program.

Recall that the independent variable, represented by x in the tables, is the natural log of the ensemble standard deviation at a grid point. The resulting dependent variable is the probability that the verification will occur in a given rank at that grid point. The distribution of probabilities from all 18 ranks makes up a verification rank histogram. PQPF for any threshold of cumulative precipitation is then determined through the weighted ranks method.

Table A1. Third-order polynomial coefficients for 1.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.1775610385	0.0638091547	-0.0011161261	-0.0017389848
2	0.0730167671	0.0221600419	-0.0007423021	-0.0007895930
3	0.0566826981	0.0154827506	-0.0006939424	-0.0006187144
4	0.0439144021	0.0101767501	-0.0006482525	-0.0004715114
5	0.0354335348	0.0066217836	-0.0006055787	-0.0003663912
6	0.0303341911	0.0044901352	-0.0005755097	-0.0003013095
7	0.0273276092	0.0031601223	-0.0005733968	-0.0002534419
8	0.0257491579	0.0021668981	-0.0006204150	-0.0001983891
9	0.0254943479	0.0012729110	-0.0007494752	-0.0001276395
10	0.0265803054	0.0003380975	-0.0009819390	-0.0000441998
11	0.0288981793	-0.0007682515	-0.0012755500	0.0000531551
12	0.0322327701	-0.0022090556	-0.0015047276	0.0001739125
13	0.0365157033	-0.0042431142	-0.0015036938	0.0003238523
14	0.0421789797	-0.0072416417	-0.0011319832	0.0004902996
15	0.0501323267	-0.0115494180	-0.0003246924	0.0006452856
16	0.0609388896	-0.0172141979	0.0008708084	0.0007704992
17	0.0739621305	-0.0238123170	0.0022945029	0.0008717093
18	0.1530469686	-0.0626406490	0.0098822733	0.0015814609

Table A2. Third-order polynomial coefficients for 2.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.1224011069	0.0462129592	0.0030028403	-0.0001408405
2	0.0629425734	0.0191124113	0.0003974715	-0.0003835922
3	0.0523940448	0.0148657563	0.0000405033	-0.0004423607
4	0.0438376060	0.0114929703	-0.0002944019	-0.0004983181
5	0.0382419099	0.0091084098	-0.0006195941	-0.0005342973
6	0.0353169303	0.0073305730	-0.0009218651	-0.0005249800
7	0.0339843883	0.0056719217	-0.0011547687	-0.0004508821
8	0.0333723629	0.0039249373	-0.0012895197	-0.0003182181
9	0.0332651455	0.0021737302	-0.0013599443	-0.0001610409
10	0.0338671813	0.0005634714	-0.0014328944	-0.0000161406
11	0.0354209811	-0.0009108059	-0.0015364163	0.0001057972
12	0.0380993826	-0.0024602339	-0.0016327575	0.0002199798
13	0.0421687845	-0.0044560885	-0.0016339727	0.0003388527
14	0.0481424036	-0.0073931691	-0.0014015055	0.0004533682
15	0.0565541426	-0.0117156145	-0.0007764539	0.0005454445
16	0.0673212360	-0.0173883491	0.0002938198	0.0006061855
17	0.0794237736	-0.0237213075	0.0016654037	0.0006375191
18	0.1432460468	-0.0524115719	0.0086540553	0.0005635236

Table A3. Third-order polynomial coefficients for 3.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.0955836770	0.0289734371	0.0020450516	0.0003949331
2	0.0555405666	0.0151473753	0.0008156506	-0.0000162624
3	0.0483806485	0.0128581072	0.0005598459	-0.0001084625
4	0.0424989663	0.0110231671	0.0002809421	-0.0002238413
5	0.0386119280	0.0097259741	-0.0000337536	-0.0003548625
6	0.0367146296	0.0087531750	-0.0003726030	-0.0004701399
7	0.0362649074	0.0078568894	-0.0006913198	-0.0005329187
8	0.0366424486	0.0069894618	-0.0009574006	-0.0005313499
9	0.0374305635	0.0061858161	-0.0011800766	-0.0004824221
10	0.0385004422	0.0052979488	-0.0013887096	-0.0004040929
11	0.0400315900	0.0039478326	-0.0016039628	-0.0002883834
12	0.0424611564	0.0017490510	-0.0018146861	-0.0001135526
13	0.0463627734	-0.0015079144	-0.0019452449	0.0001169648
14	0.0523432557	-0.0059373725	-0.0018508045	0.0003670368
15	0.0608049201	-0.0116201648	-0.0013801314	0.0005962674
16	0.0714390061	-0.0183316431	-0.0004772819	0.0007740565
17	0.0830546076	-0.0253735994	0.0007632280	0.0008792021
18	0.1373339130	-0.0557375413	0.0092312566	0.0003978274

Table A4. Third-order polynomial coefficients for 4.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.0721224212	0.0177842591	0.0001485665	0.0013324025
2	0.0485425994	0.0135232255	0.0008905945	0.0000803946
3	0.0440524983	0.0131339242	0.0010212823	-0.0002467189
4	0.0403466052	0.0128862052	0.0010597569	-0.0005606093
5	0.0380211330	0.0126441783	0.0009429220	-0.0008063599
6	0.0371593385	0.0121929324	0.0006621365	-0.0009397073
7	0.0373577870	0.0114235810	0.0002725504	-0.0009529761
8	0.0381161994	0.0104032898	-0.0001664762	-0.0008785280
9	0.0392178156	0.0092022939	-0.0006553410	-0.0007522523
10	0.0408198845	0.0077200423	-0.0012346174	-0.0005785349
11	0.0432398443	0.0057438065	-0.0018955477	-0.0003468118
12	0.0467537180	0.0030581629	-0.0025310994	-0.0000672038
13	0.0516529025	-0.0006197811	-0.0029625404	0.0002313632
14	0.0582987641	-0.0056669736	-0.0029728005	0.0005201748
15	0.0668365654	-0.0122537608	-0.0023636579	0.0007674631
16	0.0767957375	-0.0199711297	-0.0010981807	0.0009401936
17	0.0871359522	-0.0279301335	0.0006143085	0.0010306705
18	0.1335302339	-0.0632741225	0.0102681437	0.0012270399

Table A5. Third-order polynomial coefficients for 5.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.0620442287	0.0133749367	-0.0014235310	0.0011176499
2	0.0452024624	0.0122485000	0.0005906334	0.0001765333
3	0.0419024336	0.0119028293	0.0009020144	0.0000109316
4	0.0390348391	0.0119207576	0.0010443182	-0.0002139324
5	0.0371579134	0.0122754433	0.0009434673	-0.0004743604
6	0.0365631413	0.0125936585	0.0006238243	-0.0006934120
7	0.0371254879	0.0124465197	0.0001775914	-0.0007965694
8	0.0384857903	0.0116598170	-0.0003208830	-0.0007657466
9	0.0403528212	0.0103078009	-0.0008569212	-0.0006329088
10	0.0426611354	0.0085451035	-0.0014512241	-0.0004439348
11	0.0455559136	0.0064171574	-0.0021045505	-0.0002279726
12	0.0493451428	0.0036596633	-0.0027546334	0.0000244024
13	0.0544834912	-0.0003001295	-0.0032551062	0.0003460478
14	0.0614161866	-0.0059417849	-0.0033482663	0.0007160538
15	0.0701453762	-0.0131232543	-0.0027061013	0.0010048677
16	0.0799162442	-0.0210086458	-0.0011633211	0.0010632809
17	0.0895522033	-0.0285823972	0.0010698210	0.0008677699
18	0.1290551889	-0.0583959754	0.0140328680	-0.0010787003

Table A6. Third-order polynomial coefficients for 6.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0493559171	0.0138553649	0.0006293231	0.0003989841
2	0.0405808560	0.0139016745	0.0006452422	-0.0002917755
3	0.0386558363	0.0138965586	0.0007361379	-0.0004267167
4	0.0371769947	0.0137493398	0.0007565050	-0.0005284283
5	0.0364629250	0.0134039396	0.0006554589	-0.0005844533
6	0.0365484219	0.0129578944	0.0004434400	-0.0006179943
7	0.0373519124	0.0125213833	0.0001383873	-0.0006571016
8	0.0389107907	0.0119415711	-0.0002977012	-0.0006797597
9	0.0412708167	0.0108881747	-0.0008999471	-0.0006250993
10	0.0443015908	0.0092230034	-0.0016154668	-0.0004653403
11	0.0479005596	0.0070119285	-0.0023403233	-0.0002275002
12	0.0523112320	0.0041143009	-0.0029873265	0.0000592588
13	0.0580091230	-0.0000259791	-0.0034325510	0.0003935002
14	0.0652760424	-0.0060313633	-0.0034271887	0.0007615064
15	0.0739021640	-0.0139541661	-0.0026785546	0.0010928497
16	0.0831609689	-0.0229639670	-0.0010962202	0.0012897290
17	0.0920871887	-0.0317743356	0.0010774229	0.0013064577
18	0.1267366595	-0.0627153223	0.0136933623	-0.0001981168

Table A7. Third-order polynomial coefficients the 7.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0485741877	0.0113280257	-0.0002196018	0.0000371929
2	0.0391626648	0.0118360986	0.0005734400	-0.0002415247
3	0.0378097821	0.0118939480	0.0007052461	-0.0003313120
4	0.0367651668	0.0120214539	0.0008183037	-0.0004320301
5	0.0360863043	0.0123348667	0.0008881652	-0.0005428487
6	0.0358300637	0.0128713112	0.0008692285	-0.0006598437
7	0.0362028946	0.0133947778	0.0006818317	-0.0007520106
8	0.0375106177	0.0134278489	0.0002430901	-0.0007628625
9	0.0399910119	0.0125828081	-0.0004734988	-0.0006576332
10	0.0437713049	0.0107871703	-0.0014204833	-0.0004512048
11	0.0488949030	0.0081029736	-0.0025018821	-0.0001687521
12	0.0552958151	0.0044178502	-0.0035750740	0.0001997978
13	0.0627694680	-0.0005843409	-0.0044001345	0.0006670483
14	0.0709284036	-0.0071284785	-0.0045940533	0.0011598518
15	0.0791304483	-0.0149636888	-0.0037419239	0.0015022676
16	0.0865948493	-0.0232590075	-0.0017140614	0.0015397692
17	0.0927652496	-0.0310020314	0.0011419890	0.0012721558
18	0.1119168646	-0.0580615857	0.0167194188	-0.0013780608

Table A8. Third-order polynomial coefficients for 8.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0455635556	0.0056693079	-0.0007478273	0.0009771662
2	0.0373346412	0.0096259155	0.0001792018	0.0000946553
3	0.0361600367	0.0099183704	0.0004020019	0.0000183124
4	0.0355135114	0.0102689375	0.0005600100	-0.0000554556
5	0.0355084979	0.0108683701	0.0005883603	-0.0001544064
6	0.0360084980	0.0118322508	0.0004843217	-0.0003031272
7	0.0368498981	0.0130672419	0.0002993184	-0.0005053173
8	0.0381561836	0.0142024506	0.0000436280	-0.0007250544
9	0.0403568314	0.0146969718	-0.0003675114	-0.0008879469
10	0.0438667729	0.0140952019	-0.0010391473	-0.0009159956
11	0.0488242261	0.0121260346	-0.0019811240	-0.0007605351
12	0.0551741575	0.0085425217	-0.0030672975	-0.0004066152
13	0.0628094166	0.0030303061	-0.0040326606	0.0001190845
14	0.0713857845	-0.0045316176	-0.0044386150	0.0007091477
15	0.0801063866	-0.0136789807	-0.0037726649	0.0011731072
16	0.0880141693	-0.0233538652	-0.0018258635	0.0013501028
17	0.0945257285	-0.0323989571	0.0010543955	0.0012300804
18	0.1138417039	-0.0639804601	0.0176614739	-0.0009572029

Table A9. Third-order polynomial coefficients for 9.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0424368670	0.0036355507	-0.0020979281	0.0014235433
2	0.0362049905	0.0085375295	-0.0012924653	0.0009312608
3	0.0346489544	0.0097212825	-0.0007500722	0.0006384046
4	0.0330711674	0.0112113146	0.0000290314	0.0001740775
5	0.0319887391	0.0129155316	0.0008884414	-0.0004053011
6	0.0321016632	0.0144951341	0.0014932314	-0.0009271749
7	0.0338487005	0.0155575627	0.0015519259	-0.0012236110
8	0.0371432781	0.0159000573	0.0010039739	-0.0012484879
9	0.0415953188	0.0154553626	-0.0000336823	-0.0010525228
10	0.0468610945	0.0140797941	-0.0014065251	-0.0006937327
11	0.0527454050	0.0115377112	-0.0029723730	-0.0002020292
12	0.0591342152	0.0076524876	-0.0045291653	0.0003842441
13	0.0659486590	0.0022947733	-0.0056545141	0.0009567486
14	0.0730924497	-0.0047481218	-0.0056657694	0.0013288267
15	0.0803200560	-0.0135263541	-0.0039713232	0.0013288639
16	0.0871934308	-0.0233999559	-0.0006242930	0.0009308007
17	0.0932895336	-0.0331216862	0.0035979032	0.0002788926
18	0.1183754773	-0.0681979739	0.0204336038	-0.0026228035

Table A10. Third-order polynomial coefficients for 10.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0376830029	0.0060776544	-0.0027704793	0.0018335062
2	0.0338400483	0.0103913192	-0.0008175634	0.0003332639
3	0.0323880739	0.0118444595	0.0001107899	-0.0001672876
4	0.0311658889	0.0134877876	0.0009951670	-0.0006423880
5	0.0307277456	0.0152051769	0.0016328116	-0.0010451716
6	0.0314219117	0.0168552034	0.0019522877	-0.0013885700
7	0.0333406974	0.0181690327	0.0019305063	-0.0016640990
8	0.0365515905	0.0187534170	0.0014178538	-0.0017726331
9	0.0411093576	0.0182839448	0.0002101274	-0.0015899925
10	0.0468420409	0.0165845840	-0.0016983038	-0.0010690559
11	0.0534059031	0.0134734439	-0.0040345879	-0.0002501898
12	0.0606366031	0.0086735270	-0.0063634922	0.0007647831
13	0.0685906040	0.0019291212	-0.0080850454	0.0017820130
14	0.0770726207	-0.0068314620	-0.0083486443	0.0025019488
15	0.0853140905	-0.0172846486	-0.0063713219	0.0026450153
16	0.0923535343	-0.0285127697	-0.0021399055	0.0021522893
17	0.0976822809	-0.0392481707	0.0034002599	0.0012328405
18	0.1098740059	-0.0778516206	0.0289795402	-0.0036562723

Table A11. Third-order polynomial coefficients for 11.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0350030217	0.0059900724	-0.0001842269	0.0002993229
2	0.0285201947	0.0103944341	0.0016806603	-0.0006011246
3	0.0271311511	0.0117025268	0.0028503813	-0.0010590706
4	0.0261849973	0.0132282834	0.0041311586	-0.0015899875
5	0.0262942669	0.0147952202	0.0051059241	-0.0020593923
6	0.0279891629	0.0162299298	0.0052443851	-0.0022924093
7	0.0314717857	0.0174428033	0.0041271334	-0.0021512431
8	0.0365408238	0.0182952739	0.0016699979	-0.0015886856
9	0.0426863727	0.0184731657	-0.0017880192	-0.0006691181
10	0.0493419457	0.0176083867	-0.0056277889	0.0004458989
11	0.0561919581	0.0154398638	-0.0092000823	0.0015587590
12	0.0632607233	0.0116863378	-0.0118963485	0.0024934661
13	0.0706769737	0.0057578064	-0.0129948995	0.0030901108
14	0.0783828776	-0.0031462055	-0.0116700059	0.0031910357
15	0.0860438874	-0.0152211902	-0.0075018201	0.0027107562
16	0.0931685485	-0.0292407000	-0.0010872177	0.0017411331
17	0.0993457776	-0.0430050292	0.0061235590	0.0005383611
18	0.1217655311	-0.0864309794	0.0310172092	-0.0040578129

Table A12. Third-order polynomial coefficients for 12.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0236983952	0.0088951793	0.0007878787	0.0000807272
2	0.0236539482	0.0096166983	0.0038111751	-0.0009646580
3	0.0236505324	0.0107312644	0.0049065881	-0.0014268139
4	0.0240569814	0.0122805689	0.0059602535	-0.0019485215
5	0.0252522629	0.0141108014	0.0064353263	-0.0023436352
6	0.0275416431	0.0159368754	0.0058572140	-0.0024204245
7	0.0310453193	0.0174698215	0.0042213159	-0.0021535067
8	0.0357698467	0.0184731352	0.0018932138	-0.0016618364
9	0.0417553132	0.0187095797	-0.0008994095	-0.0010106372
10	0.0490044169	0.0179354383	-0.0042318410	-0.0001431047
11	0.0572054502	0.0158942201	-0.0080328230	0.0009677283
12	0.0657481855	0.0121144059	-0.0116111301	0.0021791519
13	0.0741283287	0.0057649128	-0.0137173320	0.0031980828
14	0.0821546036	-0.0038954084	-0.0131191100	0.0036996994
15	0.0897565819	-0.0165818737	-0.0094111150	0.0034980062
16	0.0967863540	-0.0305911928	-0.0034655755	0.0026843874
17	0.1030254163	-0.0437029156	0.0031167657	0.0015626913
18	0.1257664204	-0.0831615108	0.0274986050	-0.0037973364

Table A13. Third-order polynomial coefficients for 13.5-day forecast probability surface.

Rank	x^0	x^1	x^2	x^3
#	Coefficient	Coefficient	Coefficient	Coefficient
1	0.0282776712	0.0077591241	-0.0022865495	0.0008358755
2	0.0266849484	0.0084401969	0.0015530212	-0.0001285872
3	0.0269975173	0.0091122855	0.0020934130	-0.0002561455
4	0.0278198844	0.0102457811	0.0025866262	-0.0004639543
5	0.0292362388	0.0119154387	0.0029595898	-0.0007922561
6	0.0312523818	0.0139984369	0.0031103556	-0.0012143507
7	0.0339094292	0.0161923429	0.0028937115	-0.0016259997
8	0.0374540300	0.0180849327	0.0020163027	-0.0018430154
9	0.0423074561	0.0191922662	0.0000723876	-0.0016352060
10	0.0487131544	0.0190050246	-0.0031082328	-0.0008600947
11	0.0563841999	0.0170299653	-0.0070316843	0.0003715325
12	0.0645891190	0.0127615658	-0.0104362886	0.0016676986
13	0.0726517174	0.0056655392	-0.0117679359	0.0025491775
14	0.0802887240	-0.0045490389	-0.0100311113	0.0027265366
15	0.0874010221	-0.0173065328	-0.0054085893	0.0022277439
16	0.0937336913	-0.0308760562	0.0008394936	0.0013116964
17	0.0989523909	-0.0431576157	0.0071526526	0.0002848596
18	0.1133464239	-0.0735136563	0.0247928380	-0.0031555107

Table A14. Third-order polynomial coefficients for 14.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.0301418242	0.0050444712	-0.0015230981	0.0005278317
2	0.0266736757	0.0071542114	0.0033183506	-0.0009350046
3	0.0265002533	0.0086611205	0.0042128615	-0.0013093895
4	0.0268789899	0.0104609484	0.0050196842	-0.0016716002
5	0.0279590659	0.0125526419	0.0054778090	-0.0019681277
6	0.0297850351	0.0149106702	0.0053409681	-0.0021661074
7	0.0324710029	0.0172333126	0.0044623289	-0.0022194425
8	0.0363228931	0.0189135869	0.0027126932	-0.0020171241
9	0.0416568145	0.0193803005	-0.0000223084	-0.0014315598
10	0.0484838858	0.0184032504	-0.0036160663	-0.0004524118
11	0.0564444331	0.0159339381	-0.0075013371	0.0007563405
12	0.0650438847	0.0116240339	-0.0107326003	0.0019426122
13	0.0738528342	0.0046753275	-0.0122168343	0.0028595894
14	0.0824160172	-0.0055172352	-0.0110127994	0.0032684304
15	0.0900797844	-0.0183333079	-0.0068532430	0.0029918028
16	0.0961682449	-0.0317099711	-0.0006485058	0.0020951011
17	0.1004244896	-0.0433351823	0.0058709988	0.0009225532
18	0.1086968716	-0.0660521170	0.0177110985	-0.0011934936

Table A15. Third-order polynomial coefficients for 15.5-day forecast probability surface.

Rank #	x^0 Coefficient	x^1 Coefficient	x^2 Coefficient	x^3 Coefficient
1	0.0255450877	0.0077190699	-0.0007001435	0.0003332527
2	0.0253701855	0.0096644706	0.0019342288	-0.0008236392
3	0.0252538124	0.0110219695	0.0029459468	-0.0012469473
4	0.0253334757	0.0129222554	0.0042265694	-0.0018136751
5	0.0258943268	0.0151963907	0.0055999248	-0.0024910147
6	0.0273765392	0.0174487969	0.0065762099	-0.0031055361
7	0.0302561267	0.0192722665	0.0064795370	-0.0033830622
8	0.0348287315	0.0203962944	0.0047951339	-0.0030976949
9	0.0411295583	0.0205481103	0.0014459569	-0.0021810119
10	0.0490108675	0.0192867422	-0.0031337924	-0.0007346256
11	0.0581873350	0.0160899817	-0.0080814202	0.0010018359
12	0.0681737793	0.0105278979	-0.0122461765	0.0026893785
13	0.0781529705	0.0023269018	-0.0143060298	0.0039272016
14	0.0869810103	-0.0084863962	-0.0130680153	0.0043400382
15	0.0936510436	-0.0212004246	-0.0082439434	0.0037892678
16	0.0979406112	-0.0342373973	-0.0010854250	0.0025519729
17	0.1004355634	-0.0457837677	0.0062625985	0.0011619227
18	0.1064789755	-0.0727131619	0.0205988400	-0.0009176634

Appendix B: PQPF Program

This appendix is the Fortran program used to compute the PQPF by the democratic voting, uniform ranks, and weighted ranks methods. Out of the many programs written for this research, this program was included as an appendix because of its uniqueness.

```
      PROGRAM PQPF

* This program determines the probabilistic quantitative precipitation
* forecasts (PQPF) by the three different methods, each based on the
* MRF ensemble precip forecast.
* Written by Capt Tony Eckel

*~~~~~~~~~~~~~~~~~~~~~~~~~~ Variable Declarations ~~~~~~~~~~~~~~~~~~~~~~
      IMPLICIT NONE

      INTEGER           I, IE       ! index "x" grid axis, max val
      INTEGER           J, JE       ! index "y" grid axis, max val
      INTEGER           T, TE       ! index valid time, max val
      INTEGER           M, ME       ! index ensemble member, max val
      INTEGER           R, RE       ! index for RANKS, max value
      PARAMETER    (IE=25,JE=11,TE=15,ME=18,RE=18)

      REAL              COEF(TE,RE,4)! coefs 3rd order polynomials
      REAL              LIMIT(TE,2)  ! range of allowable ln(SD)
      COMMON            /A/COEF,LIMIT
      CHARACTER*30      CFILE

      CHARACTER*50      ENSPATH     ! path of ensemble datafile
      CHARACTER*50      DATEFILE
      CHARACTER*6       EFILE       ! datafile containing ensemble
      INTEGER           FS          ! IOSTAT file status
      REAL              ENS(IE,JE,TE,ME)! array for ensemble forecast
      LOGICAL           DONE

      INTEGER           EDATE
      CHARACTER*2       YR(100)     ! year of data file
      INTEGER           IYR         ! integer value for year
      INTEGER           IMN         ! integer value for month
      CHARACTER*2       MN(12)      ! month of data file
      INTEGER           IDY         ! integer value for day
      CHARACTER*2       DY(31)      ! day of data file
      INTEGER           LOM(12)     ! last day of each month

      REAL              inTOmm      ! conv factor, inches to mm
      PARAMETER         (inTOmm=25.4)
      REAL              CT(4)       ! precip cat thresholds, in mm

      REAL        DPROB(IE,JE,TE,4)! Democratic Voting PQPF
      REAL        UPROB(IE,JE,TE,4)! Uniform Ranks PQPF
      REAL        WPROB(IE,JE,TE,4)! Weighted Ranks PQPF
*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

      CT(1)=0.1 * inTOmm
```

```
      CT(2)=0.25 * inTOmm
      CT(3)=0.5 * inTOmm
      CT(4)=1.0 * inTOmm

      YR(96) = '96'
      YR(97) = '97'
      YR(98) = '98'

      DATA MN/'01','02','03','04','05','06','07','08','09',
     $        '10','11','12'/

      DATA LOM/31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31/

      DATA DY/'01','02','03','04','05','06','07','08','09','10',
     $        '11','12','13','14','15','16','17','18','19','20',
     $        '21','22','23','24','25','26','27','28','29','30',
     $        '31'/


      DO 20 T=1, 15
         CFILE = '~/Thesis/PQPF/COEF/'//DY(T)//'dayM.dat'
         OPEN (15, FILE= CFILE, ACCESS='sequential',
     $          FORM='formatted', STATUS='OLD')
        READ(15,12) LIMIT(T,1),LIMIT(T,2)
12      FORMAT(2F4.1)

        DO 18 R=1, RE
           READ(15,14) (COEF(T,R,I), I=1, 4)
14         FORMAT(4F13.10)
18      CONTINUE

        CLOSE(15)
20    CONTINUE


        DATEFILE = '~/Thesis/EnsDATA/CATALOG/fcst_M'
        OPEN(40, FILE= DATEFILE , ACCESS='sequential',
     $       FORM='formatted', STATUS='old' )

*                                  MAIN LOOP
*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
      DO WHILE (.NOT. DONE)

41      READ(40,42) EDATE
42            FORMAT(I6)

        IF (EDATE .EQ. 999999) THEN  ! End MAIN loop when reach end
           DONE = .TRUE.             !  of data marker of 999999
           GOTO 500
        ENDIF

        IF (EDATE .EQ. 0) THEN       ! 0 is an end of month marker
           GOTO 41                   !  so just skip to next EDATE
        ENDIF

        IYR = EDATE / 10000                ! Pick out int values for the
        IMN = (EDATE - IYR*10000) / 100    !  year, month, and day of
        IDY = EDATE - IYR*10000 - IMN*100  !  ensemble date to verify
```

```
*        ~~~~~~~~~~~~~~ Read in ensemble forecast data ~~~~~~~~~~~~~~~~

         EFILE = YR(IYR)//MN(IMN)//DY(IDY)

         ENSPATH = '~/Thesis/EnsDATA/Ens'//EFILE//'.12Z.dat'

         OPEN(11, FILE= ENSPATH, ACCESS='sequential',
     $        FORM='formatted', STATUS='old', IOSTAT=FS )

         IF (FS .NE. 0) THEN
            GOTO 500           ! Skip to next day if no ens file
         ENDIF                 ! available for this day

         DO 100 T = 1, TE      ! Read in entire ensemble forecast
            DO 90 M = 1, ME
             READ(11, 45) ( (ENS(I,J,T,M), I=1, IE), J=1, JE)
 45            FORMAT(25f5.1)
 90         CONTINUE
 100     CONTINUE

         CLOSE(11)

         PRINT *, ' '
         PRINT *,'------------ Forecasting for ',EFILE,' ------------'

*        ~~~~~~~ Determine PQPF by the Democratic Voting method ~~~~~~
         CALL DEMOCVOTE(DPROB,ENS,IE,JE,TE,(ME-1),CT)
         PRINT *, ':) COMPLETED Democratic Voting method'

*        ~~~~~~~ Determine PQPF by the Uniform Ranks method ~~~~~~~~~~
         CALL RANKPROB (UPROB,'U',ENS,IE,JE,TE,ME,CT)
         PRINT *, ':) COMPLETED Uniform Ranks method'

*        ~~~~~~~ Determine PQPF by the Weighted Ranks method ~~~~~~~~~~
         CALL RANKPROB (WPROB,'W',ENS,IE,JE,TE,ME,CT)
         PRINT *, ':) COMPLETED Weighted Ranks method'

         CALL DATOUT (DPROB,UPROB,WPROB,IE,JE,TE,4,EFILE)
 500  END DO

      CLOSE(40)

      END ! PROGRAM


*************************************************************************
      SUBROUTINE DEMOCVOTE (PROB,ENS,IE,JE,TE,ME,CT)

* This subroutine calculated PQPF by the democratic method

      IMPLICIT NONE

      INTEGER IE,JE,TE,ME
      REAL ENS(IE,JE,TE,ME)
      REAL PROB(IE,JE,TE,4)

      REAL CT(4)

      INTEGER CAT, I, J, T, M
```

```
      REAL COUNT


      DO 50 T = 1, TE
         DO 40 CAT = 1, 4
          DO 30 I = 1, IE
             DO 20 J = 1, JE

               COUNT = 0

               DO 10 M = 1, ME

                  IF (ENS(I,J,T,M) .GT. CT(CAT)) THEN
                     COUNT = COUNT + 1
                  ENDIF

                  IF (M .EQ. ME) THEN
                     PROB(I,J,T,CAT) = (FLOAT(COUNT) / FLOAT(ME))
     $                                       * 100.0
                  ENDIF

 10            CONTINUE

 20          CONTINUE
 30       CONTINUE
 40      CONTINUE
 50   CONTINUE

      RETURN
      END ! SUBROUTINE DEMOCVOTE


*************************************************************************
      SUBROUTINE RANKPROB (PROB,METHOD,ENS,IE,JE,TE,ME,CT)

* This subroutine calculated PQPF for the uniform ranks method or the
* weighted ranks method

* ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~  Variable Declarations ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
      IMPLICIT NONE

      CHARACTER         METHOD            ! 'U': Uniform ranks
                                          ! 'W': for weighted ranks
      INTEGER           I, IE
      INTEGER           J, JE
      INTEGER           T, TE
      INTEGER           M, ME
      INTEGER           R, RE
      PARAMETER         (RE=18)
      INTEGER           CAT               ! index category number

      REAL              PROB(IE,JE,TE,4)
      REAL              ENS(IE,JE,TE,ME)
      REAL              SUMRANKS          ! sum of prob from all ranks
      REAL              ERR               ! error for normalization
      REAL              CT(4)             ! category thresholds

      REAL              RP(18)            ! rank probabilities
      REAL              COEF(15,18,4)
```

```
      REAL              LIMIT(15,2)
      COMMON            /A/COEF,LIMIT

      REAL              FCSTPCP(17)       ! single ens fcst
      REAL              MEAN              ! mean of ens fcst
      REAL              s                 ! temp val for standard dev
      REAL              STDDEV            ! standard dev of ens fcst
      REAL              LNS               ! natural log of STDDEV
      REAL              BETA              ! beta parameter (Gumbel)
      REAL              XI                ! xi parameter (Gumbel)
      COMMON            /B/BETA,XI

      REAL              EXTRPROB          ! declare function EXTRPROB
* ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~

      DO 160 T = 1, TE

        DO 150 I = 1, IE

           DO 100 J = 1, JE

              DO 55 M = 1, ME-1
                 FCSTPCP(M) = ENS(I,J,T,M)
55            CONTINUE

              CALL SORT(FCSTPCP,17)

              MEAN = 0.0
              DO 60 M=1, 17
                MEAN = MEAN + ENS(I,J,T,M)/17.0
60            CONTINUE
              IF (MEAN .EQ. 0.0) THEN     ! Skip prob calc when all 17
                 DO 65 CAT=1, 4           !  mem fcst 0 pcp; Assign 0.0
                    PROB(I,J,T,CAT) = 0.0 !  for prob in each cat, then
65               CONTINUE                 !  advance to next grid point.
                 GOTO 100
              ENDIF

              s = 0.0

              DO 70 M=1, 17
                 s = s + (FCSTPCP(M) - MEAN)**2
70            CONTINUE
              STDDEV = SQRT (s / 16.0)

              SUMRANKS = 0.0

              DO 75 R=1, RE

                 IF (METHOD .EQ. 'W') THEN   ! make rank histogram
                    LNS = ALOG(STDDEV)

                 IF (LNS .LT. LIMIT(T,1)) THEN  ! Set LNS to min
                    LNS = LIMIT(T,1)            !  value if too low
                 ENDIF
                 IF (LNS .GT. LIMIT(T,2)) THEN  ! Set LNS to max
                    LNS = LIMIT(T,2)            !  val if too high
                 ENDIF

                 RP(R) = COEF(T,R,1) + COEF(T,R,2)*LNS +
```

```
     $                  COEF(T,R,3)*(LNS**2) + COEF(T,R,4)*(LNS**3)

                SUMRANKS = SUMRANKS + RP(R)

              ELSE
                 RP(R) = 1.0 / 18.0

              ENDIF

 75         CONTINUE

            IF (METHOD .EQ. 'W') THEN
               ERR = 1.0 - SUMRANKS
               SUMRANKS = 0.0                ! This code normalizes
               DO 77 R=1, RE                 !  the sum of the rank
                  RP(R) = RP(R) + ERR/18.0   !  prob to 1.0
                  SUMRANKS = SUMRANKS + RP(R)
 77            CONTINUE
            ENDIF

            BETA = ( STDDEV * SQRT(6.0) ) / 3.141592654   ! Gumbel
            XI = MEAN - 0.577215664 * BETA                !  parameters

*           ~~~ for the case when all thresholds < member#17 ~~~
            IF (FCSTPCP(17) .GT. CT(4)) THEN
               DO 80 CAT=1, 4
                  CALL BULKPROB(PROB(I,J,T,CAT),CT(CAT),FCSTPCP,RP)
 80            CONTINUE

*           ~~~ for the case when CAT4 threshold < member#17 ~~~
            ELSE IF (FCSTPCP(17) .GT. CT(3)) THEN
               DO 85 CAT=1, 3
                  CALL BULKPROB(PROB(I,J,T,CAT),CT(CAT),FCSTPCP,RP)
 85            CONTINUE
               PROB(I,J,T,4) = EXTRPROB('HI',CT(4),FCSTPCP(17),RP(18))

*           ~~~ for the case when CAT3&4 thresholds < member#17 ~~~
            ELSE IF (FCSTPCP(17) .GT. CT(2)) THEN
               CALL BULKPROB(PROB(I,J,T,1),CT(1),FCSTPCP,RP)
               CALL BULKPROB(PROB(I,J,T,2),CT(2),FCSTPCP,RP)
               PROB(I,J,T,3) = EXTRPROB('HI',CT(3),FCSTPCP(17),RP(18))
               PROB(I,J,T,4) = EXTRPROB('HI',CT(4),FCSTPCP(17),RP(18))

*           ~~~ for the case when CAT2,3&4 thresholds < member#17 ~~~
            ELSE IF (FCSTPCP(17) .GT. CT(1)) THEN
               CALL BULKPROB(PROB(I,J,T,1),CT(1),FCSTPCP,RP)
               DO 90 CAT=2, 4
                  PROB(I,J,T,CAT) = EXTRPROB('HI',CT(CAT),FCSTPCP(17)
     $                              ,RP(18))
 90            CONTINUE

*           ~~~ for the case when all thresholds > member#17 ~~~
            ELSE                !
               DO 95 CAT=1, 4
                  PROB(I,J,T,CAT) = EXTRPROB('HI',CT(CAT),FCSTPCP(17)
     $                              ,RP(18))
 95            CONTINUE

            ENDIF
```

```
100      CONTINUE

150    CONTINUE

160   CONTINUE

      DO 240 T = 1, TE          ! change probabilities to percentages
        DO 230 CAT = 1, 4
          DO 220 I = 1, IE
            DO 210 J = 1, JE
              PROB(I,J,T,CAT) = PROB(I,J,T,CAT) * 100.0
210         CONTINUE
220       CONTINUE
230     CONTINUE
240   CONTINUE


      RETURN
      END ! SUBROUTINE RANKPROB



*************************************************************************
      SUBROUTINE BULKPROB (P,THOLD,FCSTPCP,RP)

* This subroutine finds the bulk of the probability for one category
* for either rank type method

      IMPLICIT NONE

      REAL  P             ! (OUTPUT) probability for the category

      REAL  THOLD         ! (INPUT) the category threshold
      REAL  FCSTPCP(17)   ! (INPUT) the 17 ensemble members
      REAL  RP(18)        ! (INPUT) prob values of the 18 ranks
      REAL  EXTRPROB

      INTEGER     I       ! indexing variable


      P = 0.0
      I = 17

      DO WHILE (FCSTPCP(I) .GT. THOLD)  ! Sum prob from all verif ranks
                                        !  that exceed the threshold
         P = P + RP(I+1)
         I = I - 1

         IF (I .LT. 1) THEN
          GOTO 10
         ENDIF

      END DO

 10   IF (I .EQ. 0) THEN       ! take part of rank #1 probability
       P = P + EXTRPROB('LO',THOLD,FCSTPCP(1),RP(1))

      ELSE  ! take part of prob from rank where threshold occurs
         P = P + ( (FCSTPCP(I+1) - THOLD) /
```

```
     $             (FCSTPCP(I+1) - FCSTPCP(I)) ) * RP(I+1)
      ENDIF


      RETURN
      END ! SUBROUTINE BULKPROB


*************************************************************************
      FUNCTION GCDF (RV)

* This function returns the cumulative probability for the Gumbel
*  distribution with parameters BETA and XI for a given value of the
*  random variable.

      REAL RV                  ! value of the random variable
      REAL BETA,XI             ! parameters of the Gumbel distribution
      COMMON /B/BETA,XI


      GCDF = EXP( -1.0 * EXP( (XI - RV) / BETA ) )

      RETURN
      END ! FUNCTION GCDF


*************************************************************************
      FUNCTION EXTRPROB (HILO,THOLD,EXMEM,RP)

* This subroutine calculates the fraction of probability from either
*  extreme rank when the category threshold falls outside of the
*  ensemble members.

      CHARACTER*2 HILO         ! HI: right extreme, LO: left extreme
      REAL        THOLD        ! category threshold
      REAL        EXMEM        ! value of the extreme ensemble member
      REAL        RP           ! probability in rank #1 or #18
      REAL        GCDF         ! define GCDF as real function


      IF (HILO .EQ. 'HI') THEN
         EXTRPROB = ( (1.0 - GCDF(THOLD)) / (1.0 - GCDF(EXMEM)) ) * RP

      ELSEIF (HILO .EQ. 'LO') THEN
       EXTRPROB = ( (EXMEM - THOLD) / EXMEM ) * RP

      ENDIF

 10   RETURN
      END ! FUNCTION EXTRPROB


*************************************************************************
      SUBROUTINE SORT (ARRAY,N)

* This subprogram sorts the values within a one dimensional array from
*  least to greatest.
```

```
      IMPLICIT NONE
      INTEGER N                 ! Length of the array
      REAL ARRAY(N)             ! Array to be sorted
      INTEGER I, J              ! Index variables
      INTEGER SWAP              ! Index of array value to get swapped
      REAL MIN, TEMP

      DO 20 I=1, N-1
        MIN = ARRAY(I)
        SWAP = I
        DO 10 J=I+1, N
          IF (ARRAY(J) .LT. MIN) THEN
             MIN = ARRAY(J)
             SWAP = J
          ENDIF
10      CONTINUE

        IF (SWAP .NE. I) THEN
           TEMP = ARRAY(I)
           ARRAY(I) = MIN
           ARRAY(SWAP) = TEMP
        ENDIF

20    CONTINUE


      RETURN
      END ! SUBROUTINE SORT


*************************************************************************
      SUBROUTINE DATOUT (DP,UP,WP,XL,YL,TL,LL,FNAME)

*  This subroutine writes the PQPF to a direct access data file to be
*   used by the Gridded Data Analysis and Display System (GrADS) and
*   the program Reliability.f which measures PQPF

      IMPLICIT NONE

      INTEGER          XL,YL          ! dimensions of the grid
      INTEGER          TL             ! time dimention (# of valid points)
      INTEGER          LL             ! level dimention (# of categories)

      REAL             DP(XL,YL,TL,LL)
      REAL             UP(XL,YL,TL,LL)
      REAL             WP(XL,YL,TL,LL)

      INTEGER          X, Y, T, L     ! indicies for dimensions

      CHARACTER*6      FNAME
      INTEGER          IREC           ! index for record number
      INTEGER          ILEN           ! length of each record


      ILEN = XL * YL * 4

      OPEN (25, FILE= '~/Thesis/PQPF/FORECASTS/'//FNAME//'.12Z.dat',
     $      ACCESS='direct',FORM='unformatted', RECL=ilen)
```

```
      IREC = 0

      DO 100 T=1,TL

         DO 50 L=1, LL
            IREC = IREC +1
            WRITE (25, REC=IREC)  ( ( DP(X,Y,T,L), X=1,XL ), Y=1,YL )
50       CONTINUE

         DO 60 L=1, LL
            IREC = IREC +1
            WRITE (25, REC=IREC)  ( ( UP(X,Y,T,L), X=1,XL ), Y=1,YL )
60       CONTINUE

         DO 70 L=1, LL
            IREC = IREC +1
            WRITE (25, REC=IREC)  ( ( HP(X,Y,T,L), X=1,XL ), Y=1,YL )
70       CONTINUE

100   CONTINUE

      CLOSE (25)

      RETURN
      END ! SUBROUTINE OUTtoGRADS
```

Bibliography

Anderson, J. L., 1996: A method of producing probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518-1530

Arkansas Basin River Forecast Center (ABRFC). "What Data Do We Get?" Excerpt from an unpublished article, n. pag. WWWeb,http://info.abrbc.noaa.gov /data.html. no date

Baldwin, Michael. "Quantitative Precipitation Forecast Verification Documentation." WWWeb, http://nic.fb4.noaa.gov:8000/research/pptmethod.html, 1997

Brooks, H. E., and C. A. Doswell III, 1993: New technology and numerical weather prediction - a wasted opportunity? *Weather*, **48**, 173-177

Doran, Jeffery A. Team Leader, Weather Technology Improvement, Air Force Weather Agency, Omaha NB. Personal Correspondence. 29 August 1997

Hamill, T. M. National Center for Atmospheric Research, Research Applications Program, Boulder CO, Personal Correspondence. 22 January 1998

----, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130-141

----, 1993: *The Essence of Chaos*. University of Washington Press. 227 pp.

----, 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289-307

Mathsoft, *Mathcad User's Guide, Mathcad Plus 6.0*, Mathsoft Inc., 1995

Mullin, T., 1993: *The Nature of Chaos*. Oxford University Press. 314 pp.

National Centers for Environmental Prediction (NCEP). "Ensemble Forecasting at NCEP." Excerpt from an unpublished article, n. pag. WWWeb, http://sgi62.wwb.noaa.gov:8080/ens/info. 16 November 1995

Toth, Zoltan. Environmental Modeling Center, Camp Springs, MD. Personal communication. 9 September 1997

----, and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330

----, Y. Zhu, E. Kalnay, and S. Tracton, 1998: Probabilistic quantitative precipitation forecasts based on the NCEP global ensemble. Preprints of the 12th Conference on Numerical Weather Prediction, 11-16 January 1998, Phoenix, Arizona, in print.

Tracton, M. S., and E. Kalnay, 1993: Operational Ensemble Prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting*, **8**, 379-398

Tsonis, A. A., and J. B. Elsner, 1989: Chaos, strange attractors, and weather. *Bull. Amer. Meteor. Soc.*, **70**, 14-23

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Vita

Captain Frederick Anthony Eckel was born on 12 May 1967 in Albany, NY. He graduated from Bethlehem Central High School in 1985. In May 1989, he received a Bachelor of Science in Physics from the State University of New York at Cortland. Concurrently, he received a commission in the USAF having completed the Air Force Reserve Officer Training Corps program at Detachment 520, Cornell University.

Captain Eckel's first assignment was to complete the Basic Meteorology Program, another year of undergraduate study at Texas A & M University. In December 1990, he reported for duty as the Wing Weather Officer at McChord AFB, WA. While there, he went on numerous deployments as the weather team leader including Cairo West Air Base, Egypt, in support of Operation Restore Hope.

En route to his next assignment, he completed Squadron Officer School in residence in December 1993. He was then assigned to Yokota AFB, Japan, as the Chief of Weather Operations where he was recognized as an exceptional performer by the Air Weather Service Stan Eval team. His ideas on applying meteorological science in USAF operations have been published in the *Observer*, the magazine of the Air Weather Service. In August 1996, he was selected to enter the Air Force Institute of Technology. Following graduation, he will be assigned to the Space Warfare Center, Falcon AFB, CO, as the Air Force Weather Agency liaison.

Capt Eckel married the former Cynthia Pfaff in September of 1993.

Permanent Address: 59 Salisbury Road
Delmar, NY 12054

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1998	Master's Thesis

4. TITLE AND SUBTITLE
Calibrated Probabilistic QuantitativePrecipitation Forecasts Based on the MRF Ensemble

5. FUNDING NUMBERS

6. AUTHOR(S)
Frederick Anthony Eckel, Captain, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lt Col Michael K. Walters, Phone: (937) 255-3636 ext 4681 (DSN 785)
e-mail: mwalters@afit.af.mil
Air Force Institute of Technology
2750 P Street
Wright Patterson AFB, OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GM/98M-02

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFWA/DNXT
Mr. Jeff Doran, Phone: DSN 271-1690, email: doranj@afwa.af.mil
106 Peacekeeper Drive, Suite 2n3
Offut AFB, NE 68113-4039

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Probabilistic quantitative precipitation forecasts (PQPF) based on the medium range forecast (MRF) ensemble are currently in operational use below their full potential quality (i.e., accuracy and reliability). This unfulfilled potential is due to the MRF ensemble being adversely affected by systematic errors which arise from an imperfect model and less than ideal ensemble initial perturbations. This thesis sought to construct a calibration to account for these systematic errors and thus produce higher quality PQPF. Systematic errors were explored with the use of the verification rank histogram, which tracks the performance of the ensemble. The information in these histograms was then used in interpreting MRF ensemble forecasts to produce calibrated PQPF. While the calibration technique did noticeably improve the quality of PQPF, its usefulness was bounded by the natural predictability limits of cumulative precipitation. It was discovered that higher levels of cumulative precipitation cannot be reliably predicted in the medium range. Due to this limit of predictability, for significant levels of precipitation (high threshold), the calibration designed in this thesis was found to be useful only for short range PQPF. For low precipitation thresholds, the calibrated PQPF did prove to be of value in the medium range.

14. SUBJECT TERMS
chaos theory, ensemble weather forecasting, probabilistic weather forecasts, limits of predictability

15. NUMBER OF PAGES
145

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL